SAPNet: Segmentation-Aware Progressive Network for Perceptual Contrastive Deraining

Shen Zheng, Changjie Lu, Yuxiong Wu and Gaurav Gupta
Wenzhou-Kean University
Wenzhou, China
zhengsh, lucha, yuxiongw, ggupta@kean.edu

Figure 0. Deraining comparison at a synthetic rainy image from Rain100H (top) and a real rainy image (bottom). From left to right: Rainy, PreNet (CVPR 2019), MSPFN (CVPR 2020), MPRNet (CVPR 2021), SAPNet (ours). Compared with previous state-of-the-arts, the proposed model is superior at rain removal, edge preservation, blur suppression and color balance. Our advantage is more evident when it comes to real rainy images under complex illumination conditions.

Abstract

Deep learning algorithms have recently achieved promising deraining performances on both the natural and synthetic rainy datasets. As an essential low-level preprocessing stage, a deraining network should clear the rain streaks and preserve the fine semantic details. However, most existing methods only consider low-level image restoration. That limits their performances at high-level tasks requiring precise semantic information. To address this issue, in this paper, we present a segmentation-aware progressive network (SAPNet) based upon contrastive learning for single image deraining. We start our method with a lightweight derain network formed with progressive dilated units (PDU). The PDU can significantly expand the receptive field and characterize multi-scale rain streaks without the heavy computation on multi-scale images. A fundamental aspect of this work is an unsupervised background segmentation (UBS) network initialized with ImageNet and Gaussian weights. The UBS can faithfully preserve an image's semantic information and improve the generalization ability to unseen photos. Furthermore, we introduce a perceptual contrastive loss (PCL) and a learned perceptual image similarity loss (LPISL) to regulate model learning. By exploiting the rainy image and groundtruth as the negative and the positive sample in the VGG-16 latent space, we bridge the fine semantic details between the derained image and the groundtruth in a fully constrained manner. Comprehensive experiments on synthetic and real-world rainy images show our model surpasses top-performing methods and aids object detection and semantic segmentation with considerable efficacy. A Pytorch Implementation is available at https://github.com/ShenZheng2000/SAPNet-for-image-deraining.

1. Introduction

[1] Rain is typical weather that degrades the visibility of images and videos. Especially in heavy rain, the combination of rain streaks and accumulation has a severe adverse impact on computer vision tasks, such as image classifi-

[1]This work is supported by the research funding from Wenzhou-Kean University with grant SpF2021011.

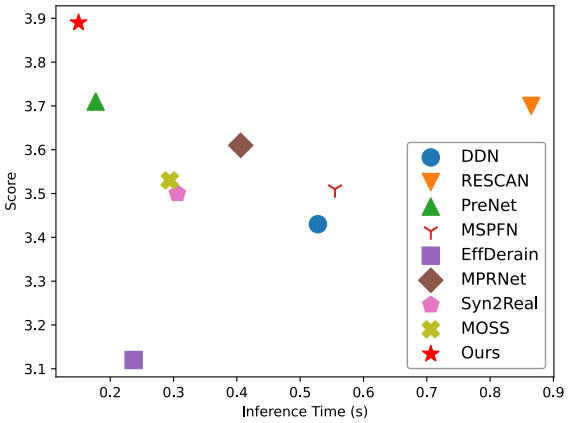


Figure 1. User study score↑ and inference time↓ comparison. The user study score (1-5) is averaged from real-rainy datasets including Rain800, SIRR and MOSS. The average run time (inference time) is calculated on images with size 512 × 512 with a single NVIDIA RTX 2080 Ti GPU.

cation, object detection, and semantic segmentation [23]. Therefore, it is crucial to remove rains and to recover the rainy images. Since 2017, deep learning deraining methods, based upon CNN [7, 48, 24, 38, 18, 36], or GAN [9, 36, 31], have attracted significant attention due to their outstanding accuracy, capacity, and flexibility.

Despite their progress in benchmark datasets, both directions focus on image quality scores like MSE/MAE and fail to consider whether their rain removal benefits high-level vision tasks such as detection and segmentation. Indeed, it has been shown by [11, 35, 23] that only considering image quality metrics does not guarantee better performance at advanced tasks. Motivated by that observation, recent models [28, 11, 59] have explored joint training to bridge the gap between low-level and high-level tasks. However, those approaches require heavy amounts of annotated images. The acquisition of those data requires tedious manual labelling, which is expensive and time-consuming. The synthetically generated labelled image also easily overfit a model, therefore compromising the generalization to real-world images.

One common way to improve the generalization ability is to transfer the knowledge from the synthetic rain domain to the real rain domain, using methods like Gaussian mixture model [44], Gaussian process [50], and self-supervised memory block [16]. Although these strategies improve real-world images that have consistent rain patterns, those deraining methods face significant performance degradation with heavy/dense rain streaks due to their failure to characterize the information from different scales and magnitudes. On the other hand, multi-scale deraining methods [53, 18, 10] require accumulating model parameters to address images resized to different scales. Consequently, the long inference time (Fig. 1) and the growing model size restrict their deployment on mobile devices or real-time deraining applications like autonomous driving and surveillance.

To address the limitations of previous researches, we propose SAPNet, a **s**egmentation-**a**ware **p**rogressive **net**work for single image deraining (Fig. 2). Due to the importance of multi-scale contextualized rain streaks information in removing heavy/dense rains, we first introduce a progressive dilated network to expand the receptive field significantly and reuse the previous recurrent stage's knowledge without additional parameters. As the semantic information is essential for task-driven deraining, we then present an unsupervised background segmentation network to preserve the semantic details during rain removal without segmentation label. Inspired by the success of contrastive learning and perceptual similarity in low-level vision tasks, we also exploit the rainy images as the negative samples to guide rain removal.

The contribution of this paper can be highlighted as four folds:

- We propose a segmentation-aware progressive network for single image deraining. To the best of our knowledge, we are first to utilize unsupervised background segmentation to aid rain removal.
- We present a novel progressive network formed with progressive dilated units (PDU). This design allows an efficient usage of multi-scale rain streak information.
- We design a new perceptual contrastive loss (PCL) and a learned perceptual image similarity loss (LPISL). With the advantage of contrastive learning and perceptual similarity, the derained image is close to the groundtruth in terms of pixel-wise difference and fine details.
- Comprehensive experiments demonstrate that our model surpasses previous state-of-the-arts qualitatively and quantitatively.

2. Related Work

In this section, we display a brief review on deep learning-based image deraining methods and contrastive learning-based image restoration approaches.

2.1. Deep Learning for Single Image Deraining

Deep Learning methods have demonstrated excellent performance in rain removal. For instance, DetailNet [7] uses a prior-based deep detail network to estimate rain streaks with negative residual information. Jorder [48] utilizes a multi-task architecture to learn binary rain streak

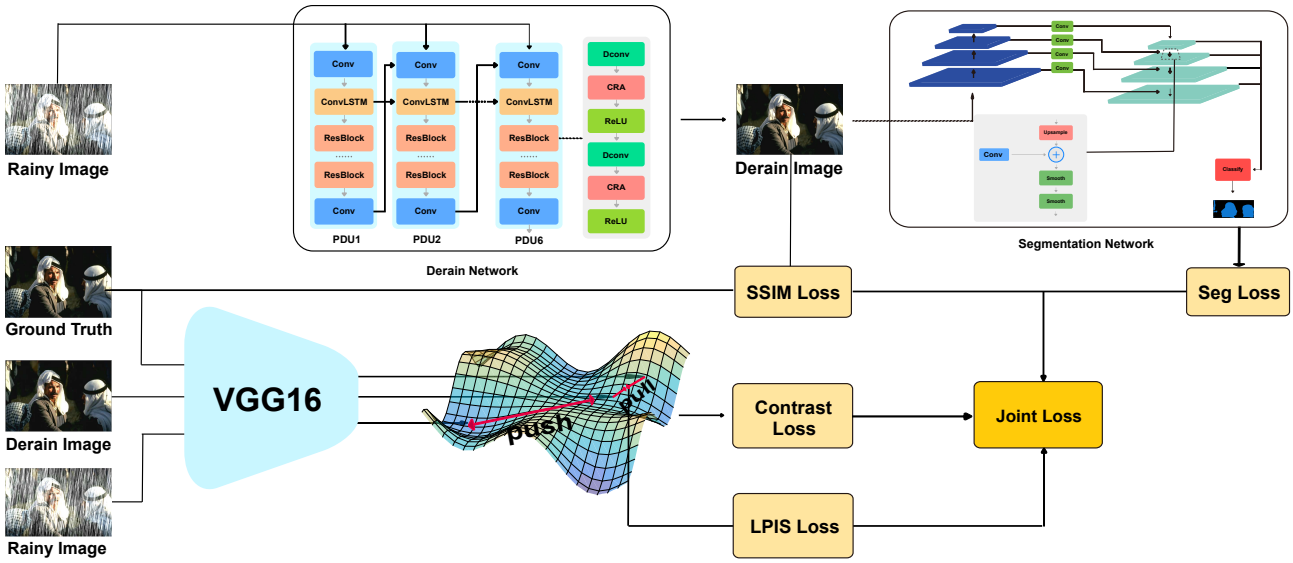


Figure 2. Model architecture for SAPNet. SAPNet joins a derain network for supervised rain removal, a segmentation network for unsupervised background segmentation, and a VGG-16 network for perceptual contrast. The rainy image first enters the derain network for rain removal, using the groundtruth as the reference to obtain the negative ssim loss. Once the deraining process is finished, the segmentation network will consume the derained image to calculate the segmentation loss. Meanwhile, the perceptual contrastive loss and the learned perceptual image similarity loss will be computed on the VGG16 latent space using the rainy image, the derained image, and the groundtruth. Finally, the jointed loss will update the derain network during training. Here we use 'DConv' for dilated convolution and 'CRA' for the channel residual attention block in Fig. 3

maps, heavy rain streaks appearance, and the clean background in one go.

Recurrent networks have been employed to construct more efficient image deraining models. For example, RESCAN [24] leverages a recurrent neural network with squeeze-and-excitation blocks for rain removal. PreNet [38] recursively unfolds a shallow residual network to process the input and intermediate layers progressively. ID-CGAN [36] proposed an attention-based GAN [9] for attending raindrops and its neighbor backgrounds.

Recent deraining methods [49, 18, 52] have begun to corporate multi-scale learning to exploit rain streaks of different sizes and directions. For instance, MSPFN [18] utilizes a multi-scale pyramid architecture to supervise the fine fusion of rain streaks information.

Different from previous approaches, we utilize dilated convolution [51] to expand the receptive field. In this way, we obtain rain streaks of different scales within one recurrent unit without compromising the computational efficiency. We also exploit unsupervised semantic segmentation to restore the background semantic details during intensive rain removal.

2.2. Contrastive Learning for Image Restoration

Contrastive Learning have made notable progress in self-supervised representation learning [3, 12, 14]. The goal of contrastive learning is to pull an anchored sample near to the positive sample and, meanwhile, push that anchored sample away from the negative sample in the given latent space. Previous contrastive learning often targets high-level vision tasks like image classification and object detection.

Recently, contrastive learning has been utilized in low-level vision tasks. For example, [32] has shown that contrastive learning can boost the performance of unpaired image-to-image translation. Contrastive Learning is also applied in image dehazing [46] with a pixel-wise L1 loss and in image super-resolution [43] with self-supervised knowledge distillation.

Unlike previous contrastive learning methods, this work applies contrastive learning to single image deraining for the first time. To better reserve the fine details in a photo during rain removal, we take perceptual similarity into account and present a new perceptual contrastive loss.

3. Methodology

In this section, we present the proposed SAPNet by analyzing the building components, network architecture, and loss functions.

3.1. Channel Residual Attention Block

Building blocks are essential for rain removal because they determine a model's ability to characterize the rain streak patterns. Recently, state-of-the-arts deraining methods [24, 36, 18, 52] have begun to incorporate the attention

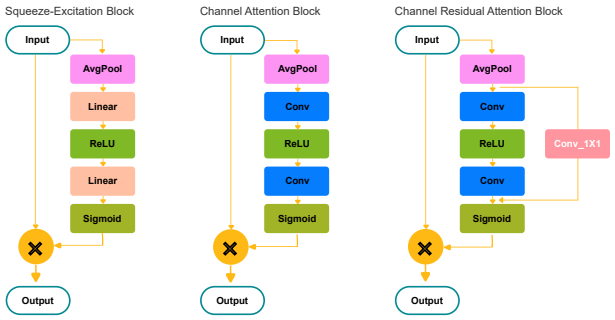


Figure 3. Different attention blocks for image deraining. From left to right: squeeze-excitation (SE) block, channel attention (CA) block, and the proposed channel residual attention (CRA) block.

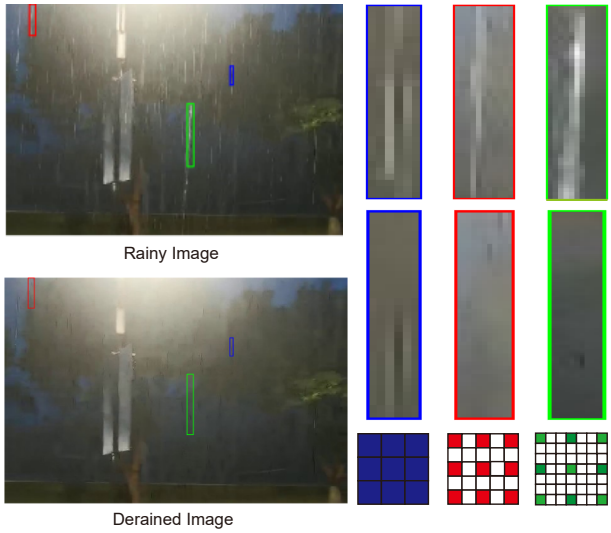


Figure 4. Visual Illustration of Progressive Dilation. We use blue, red, green bounding boxes to highlight rain streaks of diverse shapes and thickness. In the proposed network, each progressive dilated unit utilize convolutions with different dilation rate to capture and clear multi-scale contextualized rain streaks.

mechanism to boost deraining performances. In Fig. 3, we display three effective blocks for image deraining, including squeeze-excitation (SE) block [15], channel attention (CA) block [45] and our proposed channel residual attention (CRA) block. Compared with SE and CA, our skip connection from the pooling layer to the sigmoid activation function allows a more efficient feature fusion and gradient flow during the model training. The ablation study will analyze the superiority of CRA qualitatively and quantitatively.

3.2. Progressive Dilated Unit

A basic neural network like [7] cannot characterize heavy/dense rain streaks, as shown by [38, 52].

Inspired by the success of progressive networks [24, 38] for image deraining and the efficiency of dilated convolution for multi-scale information [1, 2] , we design a progressive dilated unit (PDU), which uses dilated convolution to exploit the multi-scale contextualized rain streaks information (See Fig. 4). Our PDU contains four parts, a leading convolutional block for consuming the input rainy image, five proposed residual blocks for feature extraction and an ending convolution block for yielding the derained image. The dilation rate of the residual blocks are 1, 2, 4, 8, 16, respectively.

Each proposed residual block includes, in a single repetition, a convolution layer, a channel attention block, and a ReLU [30] activation layer. The channel number of the convolution layers is 32, and the kernel size is 3. Besides, the reduction factor of the CRA is 16. Finally, we have an output convolution layer that reduces the channel number from 32 to 3.

3.3. Derain Network

Our derain network consists of 6 recurrent stages. Each stage corresponds to a progressive dilated unit (PDU) which has shared parameters with others. The inference of our network is:

$$\begin{aligned} \mathbf{s}^t &= f_{\text{rec}}\left(\mathbf{s}^{t-1}, f_{\text{in}}\left(\mathbf{x}^{t-1}, \mathbf{y}\right)\right) \\ \mathbf{x}^t &= f_{\text{out}}\left(f_{\text{res}}\left(\mathbf{s}^t\right)\right) \end{aligned} \tag{1}$$

where $\mathbf{y}$ is the rainy image. where f_{in} and f_{out} is the convolutional block for receiving the input and outputting the results, respectively. f_{rec} is the recurrent operations. f_{res} is the residual blocks. $\mathbf{s}^t$ is the recurrent state at stage t. $\mathbf{x}^t$ is the derained image at stage t. Note that we combine the rainy image and the derained image from previous recurrent unit as the input for the next recurrent unit. That strategy is shown by [36, 38] to boost deraining performance.

For the recurrent calculations, we leverage convolutional LSTM [47] for a more consistent cross-stage interaction. It can be formulated as:

$$\begin{aligned} i_t &= \sigma\left(W_{xi} * x_t + W_{hi} * h_{t-1} + W_{ci} \circ c_{t-1} + b_i\right) \\ f_t &= \sigma\left(W_{xf} * x_t + W_{hf} * h_{t-1} + W_{cf} \circ c_{t-1} + b_f\right) \\ c_t &= f_t \circ c_{t-1} + i_t \circ \tanh\left(W_{xc} * x_t + W_{hc} * h_{t-1} + b_c\right) \\ o_t &= \sigma\left(W_{xo} * x_t + W_{ho} * h_{t-1} + W_{co} \circ c_t + b_o\right) \\ h_t &= o_t \circ \tanh\left(c_t\right) \end{aligned} \tag{2}$$

where i_t is the input gate, f_t is the forget gate, o_t is the output gate c_t and is the cell state, $\circ$ denotes the element-wise product, and $*$ denotes the convolution operation.

3.4. Segmentation Network

Semantic segmentation has been useful for low-level vision tasks such as image denoising [28, 27], image deblurring [39] and image deraining [54]. Inspired by this, we design an unsupervised background segmentation network that performs semantic segmentation on the derained image. Similar to [28, 40], we freeze the parameters of the entire segmentation network during training.

Name	Category	Test Samples
Rain12	Synthetic	12
Rain100L	Synthetic	100
Rain100H	Synthetic	100
Rain800	Real	50
SIRR	Real	147
MOSS	Real	48
COCO150	Synthetic	150
CityScape150	Synthetic	150

Table 1. Dataset Description

Motivated by the success of the feature pyramid network [25] in utilizing multi-scale contextual information, which is essential for image deraining, our segmentation network uses an FPN backbone which consists of an encoder-decoder framework with lateral connections embedded with 1×1 convolution layer. Our encoder (bottom-up pathway) use ResNet-101 [13] pretrained on ImageNet [5], whereas our decoder (top-down pathway) is initialized with Gaussian weight with zero mean and a standard deviation of 0.05. Besides, both the encoder and the decoder have four convolution blocks.

For each decoder stairs, the output image is bilinearly upsampled and concatenated with the lateral results. Two smooth layers of 3×3 convolution are designed for better perceptual quality after each concatenation. Finally, all stairs' image in the decoder is concatenated. The concatenated output's channel number is reduced from 512 to n, leading to a pixel-wise classification task with n-class. Empirically, we set n to 21 because the benchmark image classification dataset PASCAL-VOC 2012 [6] have 20 significant object classes and another background class.

3.5. Loss Function

Negative SSIM Loss Most single image deraining tasks use L2 loss for training. As shown by [8, 38, 41, 18], the L2 loss produces over-smoothed backgrounds and ghost artifacts, which is detrimental to the semantic information. As an alternative, we adopt negative SSIM loss to focus on luminance, contrast, and structure. The negative SSIM loss is:

$$\mathcal{L}_{\text{ssim}} = -\,\text{SSIM}\left(\mathbf{x}^D, \mathbf{x}^G\right) \tag{3}$$

where $\mathbf{x}^D$, $\mathbf{x}^R$, and $\mathbf{x}^G$ represents the derained image, the rainy image, and the groundtruth, respectively. is the rainy image.

Segmentation Loss For unsupervised semantic segmentation of UBS, We utilize focal loss [25] to address the imbalance for rain streaks of different directions and magnitudes. Since the segmentation label is unavailable, we minimize the average of the cost function to make an overall more 'confident' prediction. The segmentation loss is:

$$\mathcal{L}_{\text{seg}} = \frac{1}{HW} \sum_{1 \le i \le H, 1 \le j \le W} -\alpha \left(1 - p_{i,j}\right)^{\gamma} \log p_{i,j} \tag{4}$$

where H, W is the height and the width of the image. $p_{i,j}$ is the model's estimated probability for the class with a specific pixel-wise class probability in segmentation. Here we set α equals to 1 and γ as 2.

Perceptual Contrastive Loss A simple contrastive loss is usually based upon L1 loss [32, 46, 43]. However, it is shown by [19] that simple pixel-wise loss (L1/L2 loss) fails to reserve fine details and textures during image processing. Inspired by the success of perceptual loss [19] in low-level vision tasks like image-to-image traslation [17], image super-resolution [22], and image deblurring [21], we inject perceptual loss into contrastive loss. The proposed perceptual contrastive loss is:

$$\mathcal{L}_{\text{pcl}} = \sum_{i=1}^{n} \omega_i \cdot \frac{L1\left(V_i(\mathbf{x}^D), V_i(\mathbf{x}^G)\right)}{L1\left(V_i(\mathbf{x}^D), V_i(\mathbf{x}^R)\right)} \tag{5}$$

where V_i represents the i^{th} extracted layer in from VGG-16. ω_i represents the weight coefficient to balance between shallow and deep layer features.

Learned Perceptual Image Similarity Loss Learned Perceptual Image Patch Similarity (LPIPS) is first proposed in [55] to evaluate the perceptual similarity between the distorted image and the groundtruth. In this paper, we use the resized whole image rather than the cropped image patches proposed by the original paper. There are two reasons for this change. First, operating on the whole image helps restore the high-level semantic information crucial for detection and segmentation [58]. Second, perceptual similarity on the entire image explores non-local information [42], thereby complementing the convolution operations which can only process one local region at a time. We name our loss Learned Perceptual Image Similarity Loss (LPISL). The formulation is as below:

$$\mathcal{L}_{\text{lpisl}} = \sum_{i=1}^{n} \frac{1}{H_i W_i} \sum_{h,w} \left\| \theta_i \odot \left(V_i(\mathbf{x}^D) - V_i(\mathbf{x}^G)\right) \right\|_2^2 \tag{6}$$

where θ_i represents the cosine distance calculation.

Total Loss Our total loss for SAPNet is:

$$\mathcal{L} = \lambda_1 \times \mathcal{L}_{\text{ssim}} + \lambda_2 \times \mathcal{L}_{\text{seg}} + \lambda_3 \times \mathcal{L}_{\text{pcl}} + \lambda_4 \times \mathcal{L}_{\text{lpisl}} \tag{7}$$

Here we set λ_1 to 1, λ_2, λ_3 and λ_4 to 0.1

Figure 5. Visual ablation of attention blocks. Top row: original image. Bottom row: cropped image. From left to right: Model-Conv, Model-SE, Model-CA, Model-CRA.

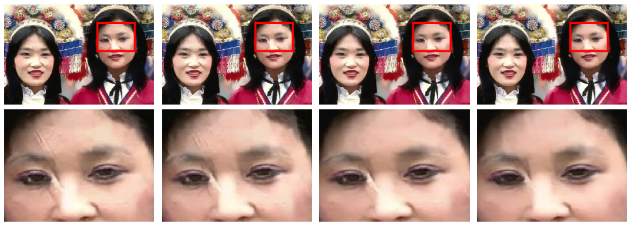


Figure 6. Visual ablation of model components. Top row: original image from Model 1 to 4. Bottom row: cropped image from Model 1 to 4.

Figure 7. Visual ablation of contrastive losses. From left to right, PreNet (No-CL), SAPNet (No-CL), SAPNet (L1-CL), SAPNet (PCL)

	Model-SE	Model-CA	Model-CRA
PSNR	28.74	28.89	**29.46**
SSIM	0.890	0.892	**0.897**
RT	0.166	**0.149**	0.150

Table 2. Ablation on attention blocks in terms of PSNR↑, SSIM↑ and Run Time↓

	Contrastive Loss			Prop. of Train Images		
Metric	No-CL	L1-CL	PCL	40 %	60 %	100 %
PSNR	28.96	26.84	**29.46**	26.49	27.37	**29.46**
SSIM	0.888	0.853	**0.897**	0.853	0.866	**0.897**

Table 3. Ablation result for SAPNet with different contrastive lose and limited training images

	M1	M2	M3	M4	M5	Ours
CRA	✓	✓	✓	✓	✓	✓
UBS		✓	✓	✓	✓	✓
PCL			✓	✓	✓	✓
Dilation				✓	✓	✓
Decay					✓	✓
LPISL						✓
PSNR	27.94	28.34	28.56	28.93	29.36	**29.46**
SSIM	0.882	0.886	0.887	0.891	0.896	**0.897**

Table 4. Ablation result for SAPNet with different model (M) components.

4. Experiments

4.1. Implementation Details

The proposed model is trained using Pytorch [33] with one Tesla V100 GPU. The training dataset RainTrain100H [48] contains 1800 pairs of synthetic rainy images and the corresponding ground truth. The proposed model uses Adam [20] optimizer for 100 epochs with an initial learning rate of 0.001 and a batch size of 12. The learning rate is reduced by 80 % on epochs 30, 50, and 80, respectively. It takes around 20 hours for the model to converge.

4.2. Datasets

We utilize benchmark synthetic and real-world rainy datasets for comparisons. To investigate task-driven image deraining, we additionally chose 300 images in total from Microsoft COCO [26] and CityScape [4]. We synthesize rain for them and name the dataset COCO150 and CityScape150, respectively. Dataset details is at Table 1.

4.3. Metrics

Several benchmark metrics have been adopted for assessment. For the synthetic dataset, we use scikit-learn [34] as a unified library[2] for PSNR and SSIM. For the real-world dataset, we use non-reference metrics, including UNIQUE [56] and BRISQUE [29]. For the task-driven parts, we use mean average precision (mAP) for object detection, mean pixel accuracy (mPA) and mean intersection over union (mIOU) for semantic segmentation. The method with the best and the second-best score is in **bold** and underline, respectively.

4.4. Baselines

We compare the proposed method with recent state-of-the-arts. The superivsed method for comparison includes DDN [7], RESCAN [24], PreNet [38], MSPFN [18], MPRNet [52], and EffDerain [10]. The unsupervised method includes Syn2Real [50] and MOSS [16]. To ensure a fair comparison, all supervised methods for comparison is trained on RainTrain100H without data augmentation, using their publicly available codes.

4.5. Ablation Study

We conduct comprehensive ablation studies to investigate the contributions of each component to SAPNet's rain removal performance. The ablation studies are evaluated on synthetic rainy datasets due to the requirements of PSNR and SSIM.

[2]Most papers use Matlab for computing their PSNR and SSIM. We find that, holding everything else fixed, the PSNR calculated from sklearn will be 1-2 db lower than the result from Matlab.

Methods	Rain12	Rain100L	Rain100H
Rainy	28.82/0.836	25.52/0.825	12.13/0.349
DDN	28.89/0.897	26.25/0.856	12.65/0.420
RESCAN	33.60/0.953	31.76/0.946	27.43/0.841
PreNet	34.79/0.964	**36.09**/0.972	28.06/0.884
Syn2Real	28.06/0.893	24.24/0.871	15.18/0.397
MSPFN	34.17/0.945	30.55/0.915	26.29/0.798
MOSS	28.82/0.835	27.27/0.885	16.82/0.487
EffDerain	28.11/0.836	25.72/0.800	14.82/0.439
MPRNet	**36.53**/0.963	34.73/0.959	28.52/0.872
Ours	35.50/**0.968**	34.77/**0.973**	**29.46/0.897**

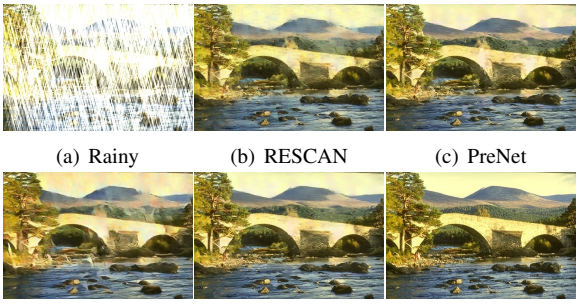
Table 5. PSNR ↑ and SSIM ↑ comparison on Rain12, Rain100L and Rain100H

Methods	Rain800	SIRR	MOSS
Rainy	0.755/26.63	0.672/29.13	0.786/26.47
DDN	0.741/**18.12**	0.670/25.46	0.790/19.92
RESCAN	0.761/21.54	0.671/25.67	0.794/19.02
PreNet	0.762/20.08	0.674/24.17	0.797/18.26
Syn2Real	0.750/20.04	0.689/24.11	0.783/17.96
MSPFN	0.749/22.17	0.657/20.71	0.732/22.64
MOSS	0.743/22.05	0.691/29.06	0.788/24.45
EffDerain	0.737/31.86	0.679/39.33	0.773/38.10
MPRNet	0.754/21.57	**0.697**/28.48	0.797/24.22
Ours	**0.767**/22.21	0.696/**20.68**	**0.798/17.88**

Table 6. UNIQUE ↑ / BRISQUE ↓ comparison on Rain800, SIRR, and MOSS

Metrics	Rainy	DDN	RESCAN	PreNet	EffDerain	Syn2Real	MOSS	**Ours**	GT
mAP (%)	52.1	65.1	78.5	81.0	68.2	55.4	73.2	**82.2**	85.4
mPA (%)	65.3	66.4	70.3	73.8	67.3	59.9	76.6	**77.2**	78.8
mIOU (%)	50.7	53.6	57.3	56.3	56.7	49.9	60.1	**62.2**	66.7

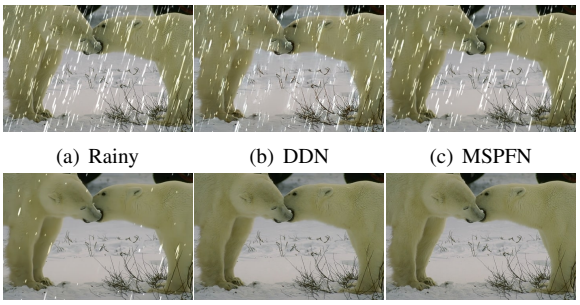
Table 7. mAP↑, mPA↑ and mIOU↑ comparison



(a) Rainy (b) RESCAN (c) PreNet

(d) MSPFN (e) Ours (f) GT

Figure 8. Visual comparison at Rain100H



(a) Rainy (b) DDN (c) MSPFN

(d) MPRNet (e) Ours (f) GT

Figure 9. Visual comparison at Rain100L

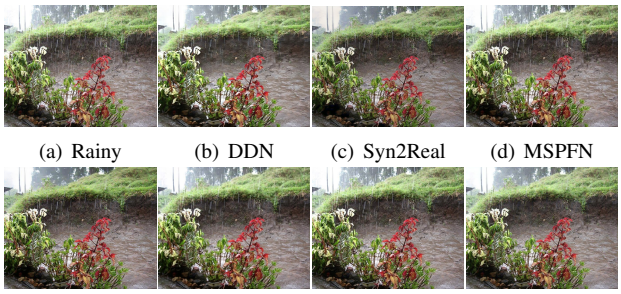
Ablation of Attention Blocks The first ablation study aims to demonstrate the superiority of the proposed channel residual attention (CRA) block compared with the squeeze-excitation (SE) block and channel attention (CA) block. Table 2 shows the PSNR and SSIM for SAPNet with different attention blocks. We note that SAPNet-CRA has the best PSNR and SSIM with an efficient inference time. Fig. 5 displays the corresponding visual comparison. We can see that both SAPNet-Conv and SAPNet-SE fail to clear rain streaks in the sky. Although SAPNet-CA effectively removes large rain streaks, it over-smooth the backgrounds. In comparison, SAPNet-CRA preserves the most background textures and has the most pleasant looking.

Ablation of Model Components The second ablation study examines the effectiveness of different model components for SAPNet. Table 4 shows different versions of SAPNet, where we sequentially add channel residual attention (CRA), unsupervised background segmentation (UBS), perceptual contrastive loss (PCL), dilation, learning rate decay and learned perceptual image similarity loss (LPISL). We notice that each component contributes to better rain removal (i.e., better PSNR and SSIM). The visual comparison in Fig. 6 also demonstrates that the proposed modules help rain removal and facial details preservation.

Ablation of Contrastive Losses The third ablation study investigates the effectiveness of the proposed perceptual contrastive loss (PCL). Table 3 compares SAPNet's performance with no contrastive loss, L1 contrastive loss, and perceptual contrastive loss. We can see that L1 contrastive loss significantly degrade the rain removal performances, whereas the perceptual contrastive loss substantially improves the performances. We also make a visual comparison in Fig. 7, with PreNet as an additional reference. It shows that SAPNet with no contrastive loss or with L1 contrastive loss fails in large and long rain streaks. In comparison, SAPNet with the proposed PCL successfully remove different types of rain streaks.

4.6. Comparison on Synthetic Rainy Dataset

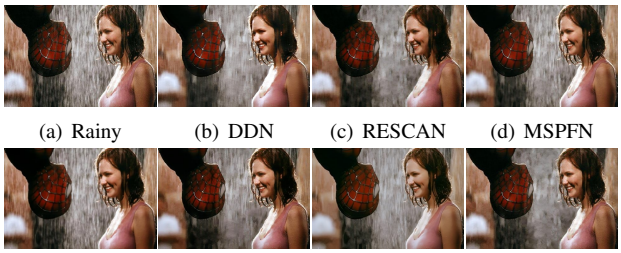
We make a quantitative comparison for synthetic rainy datasets in Table 5. It can be seen that SAPNet has the best SSIM for all, and the second-best PSNR for Rain12 and Rain100L. For the most challenging Rain100H, SAPNet has the best PSNR and SSIM. We also conduct a visual comparison on synthetic rainy images. Fig 8 (Rain100H) shows that RESCAN, PreNet, and MSPFN clear most heavy rain streaks but leave significant grey marks on the background sky. In comparison, SAPNet has the best rain removal performance and is closest to the groundtruth. Fig 9



(a) Rainy (b) DDN (c) Syn2Real (d) MSPFN

(e) MOSS (f) EffDerain (g) MPRNet (h) Ours

Figure 10. Visual comparison at Rain800



(a) Rainy (b) DDN (c) RESCAN (d) MSPFN

(e) EffDerain (f) MOSS (g) MPRNet (h) Ours

Figure 11. Visual comparison at SIRR

(Rain100L) shows that DDN and MSPFN fail to clear the long rain streaks and that MPRNet's local details are unpromising. In contrast, SAPNet's derained image is almost on par with the groundtruth.

4.7. Comparison on Real-World Rainy Images

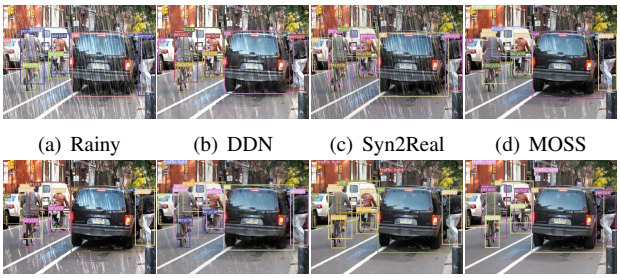
We make a quantitative comparison for real-world rainy datasets in Table 6. It shows that SAPNet has the best BRISQUE for SIRR and the best UNIQUE for Rain800. For the recently proposed MOSS dataset, SAPNet outperforms all competing models in terms of UNIQUE and BRISQUE. We also conduct qualitative comparisons on Rain800 (Fig. 10) and SIRR (Fig. 11). It shows that other methods (1) fail to clear the rain streaks (2) introduce blur and under/over-exposure. In contrast, SAPNet maintains the best brightness and exposure while removing diverse types of rain streaks effectively.

4.8. Model Efficiency

Real-time deraining on mobile devices demands an affordable model size with a fast inference speed. It is essential to investigate the model efficiency regarding the number of parameters and the inference time (Fig. 1). More details will be in the supplementary material.

4.9. Detection and Segmentation

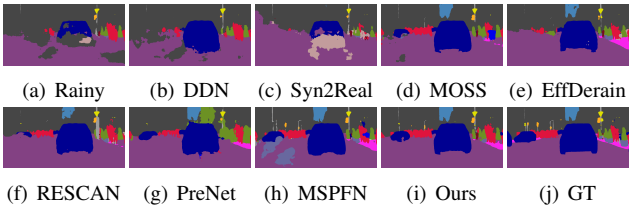
To investigate the contribution of deraining models to high-level vision tasks, we use Yolov3 [37] for object detection on COCO150 and PSPNet [57] for semantic segmentation on CityScape 150.



(a) Rainy (b) DDN (c) Syn2Real (d) MOSS

(e) EffDerain (f) RESCAN (g) Ours (h) GT

Figure 12. Object detection result at COCO150



(a) Rainy (b) DDN (c) Syn2Real (d) MOSS (e) EffDerain

(f) RESCAN (g) PreNet (h) MSPFN (i) Ours (j) GT

Figure 13. Semantic segmentation result at CityScape150

Fig. 12 reveals that the competing models have limited ability to conduct rain removal for object detection. For example, DDN, Syn2Real, and EffDerain fail to detect the traffic lights, which could be disastrous for automatic pilots. In contrast, SAPNet removes most rain streaks and helps bridge the detection to the groundtruth.

Fig. 13 displays that the deraining method for comparison also has limited contribution to semantic segmentation. For instance, DDN, Syn2Real, EffDerain, and RESCAN miss the left car in the segmentation map. In comparison, SAPNet has the most accurate segmentation and is closest to the groundtruth. The quantitative comparisons in Table 7 further demonstrates that SAPNet has the best performance in both object detection and semantic segmentation.

5. Conclusion

This paper presented a segmentation-aware progressive network for image deraining. Firstly, we designed a progressive dilated unit (PDU) to utilize the multi-scale rain streaks information. Secondly, we proposed perceptual contrastive loss (PCL) and learned perceptual image similarity loss (LPISL) to bridge the derained image to the groundtruth in terms of pixel-wise and perceptual-level differences. Finally, we leveraged unsupervised background segmentation (UBS) to reserve the semantic information during exhaustive rain removal. Extensive experiments demonstrate the effectiveness of the proposed method. Our future work will explore detection-driven deraining and investigate rain removal at sub-optimal illumination.

References

[1] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[2] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[4] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[6] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[7] 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 Conference on Computer Vision and Pattern Recognition*, pages 3855–3863, 2017.

[8] Xueyang Fu, Borong Liang, Yue Huang, Xinghao Ding, and John Paisley. Lightweight pyramid networks for image deraining. *IEEE transactions on neural networks and learning systems*, 31(6):1794–1807, 2019.

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[10] Qing Guo, Jingyang Sun, Felix Juefei-Xu, Lei Ma, Xiaofei Xie, Wei Feng, and Yang Liu. Efficientderain: Learning pixel-wise dilation filtering for high-efficiency single-image deraining. *arXiv preprint arXiv:2009.09238*, 2020.

[11] Muhammad Haris, Greg Shakhnarovich, and Norimichi Ukita. Task-drivensuper resolution: Object detection in low-resolution images. *arXiv preprint arXiv:1803.11316*, 2018.

[12] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[14] Olivier Henaff. Data-efficient image recognition with contrastive predictive coding. In *International Conference on Machine Learning*, pages 4182–4192. PMLR, 2020.

[15] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[16] 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.

[17] 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.

[18] Kui Jiang, Zhongyuan Wang, Peng Yi, Chen Chen, Baojin Huang, Yimin Luo, Jiayi Ma, and Junjun Jiang. Multi-scale progressive fusion network for single image deraining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8346–8355, 2020.

[19] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

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

[21] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiří Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8183–8192, 2018.

[22] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

[23] Siyuan Li, Iago Breno Araujo, Wenqi Ren, Zhangyang Wang, Eric K Tokuda, Roberto Hirata Junior, Roberto Cesar-Junior, Jiawan Zhang, Xiaojie Guo, and Xiaochun Cao. Single image deraining: A comprehensive benchmark analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3838–3847, 2019.

[24] Xia Li, Jianlong Wu, Zhouchen Lin, Hong Liu, and Hongbin Zha. Recurrent squeeze-and-excitation context aggregation net for single image deraining. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 254–269, 2018.

[25] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In

European conference on computer vision, pages 740–755. Springer, 2014.

[27] Ding Liu, Bihan Wen, Jianbo Jiao, Xianming Liu, Zhangyang Wang, and Thomas S Huang. Connecting image denoising and high-level vision tasks via deep learning. *IEEE Transactions on Image Processing*, 29:3695–3706, 2020.

[28] Ding Liu, Bihan Wen, Xianming Liu, Zhangyang Wang, and Thomas S Huang. When image denoising meets high-level vision tasks: A deep learning approach. *arXiv preprint arXiv:1706.04284*, 2017.

[29] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012.

[30] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Icml*, 2010.

[31] Jinshan Pan, Jiangxin Dong, Yang Liu, Jiawei Zhang, Jimmy Ren, Jinhui Tang, Yu-Wing Tai, and Ming-Hsuan Yang. Physics-based generative adversarial models for image restoration and beyond. *IEEE transactions on pattern analysis and machine intelligence*, 43(7):2449–2462, 2020.

[32] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *European Conference on Computer Vision*, pages 319–345. Springer, 2020.

[33] 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*, 32:8026–8037, 2019.

[34] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[35] Yanting Pei, Yaping Huang, Qi Zou, Yuhang Lu, and Song Wang. Does haze removal help cnn-based image classification? In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 682–697, 2018.

[36] Rui Qian, Robby T Tan, Wenhan Yang, Jiajun Su, and Jiaying Liu. Attentive generative adversarial network for raindrop removal from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2482–2491, 2018.

[37] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[38] Dongwei Ren, Wangmeng Zuo, Qinghua Hu, Pengfei Zhu, and Deyu Meng. Progressive image deraining networks: A better and simpler baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3937–3946, 2019.

[39] Wenqi Ren, Jinshan Pan, Xiaochun Cao, and Ming-Hsuan Yang. Video deblurring via semantic segmentation and pixel-wise non-linear kernel. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1077–1085, 2017.

[40] Sicheng Wang, Bihan Wen, Junru Wu, Dacheng Tao, and Zhangyang Wang. Segmentation-aware image denoising without knowing true segmentation. *arXiv preprint arXiv:1905.08965*, 2019.

[41] Tianyu Wang, Xin Yang, Ke Xu, Shaozhe Chen, Qiang Zhang, and Rynson WH Lau. 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*, pages 12270–12279, 2019.

[42] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[43] Yanbo Wang, Shaohui Lin, Yanyun Qu, Haiyan Wu, Zhizhong Zhang, Yuan Xie, and Angela Yao. Towards compact single image super-resolution via contrastive self-distillation. *arXiv preprint arXiv:2105.11683*, 2021.

[44] Wei Wei, Deyu Meng, Qian Zhao, Zongben Xu, and Ying Wu. Semi-supervised transfer learning for image rain removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3877–3886, 2019.

[45] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[46] 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.

[47] SHI Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems*, pages 802–810, 2015.

[48] Wenhan Yang, Robby T Tan, Jiashi Feng, Jiaying Liu, Zongming Guo, and Shuicheng Yan. Deep joint rain detection and removal from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1357–1366, 2017.

[49] Rajeev Yasarla and Vishal M Patel. 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*, pages 8405–8414, 2019.

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

[51] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

[52] 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.

[53] He Zhang, Vishwanath Sindagi, and Vishal M Patel. Image de-raining using a conditional generative adversarial network. *IEEE transactions on circuits and systems for video technology*, 30(11):3943–3956, 2019.

[54] Kaihao Zhang, Wenhan Luo, Wenqi Ren, Jingwen Wang, Fang Zhao, Lin Ma, and Hongdong Li. Beyond monocular deraining: Stereo image deraining via semantic understanding. In *European Conference on Computer Vision*, pages 71–89. Springer, 2020.

[55] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[56] Weixia Zhang, Kede Ma, Guangtao Zhai, and Xiaokang Yang. Uncertainty-aware blind image quality assessment in the laboratory and wild. *IEEE Transactions on Image Processing*, 30:3474–3486, 2021.

[57] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.

[58] Shen Zheng and Gaurav Gupta. Semantic-guided zero-shot learning for low-light image/video enhancement. *arXiv preprint arXiv:2110.00970*, 2021.

[59] Shen Zheng, Yuxiong Wu, Shiyu Jiang, Changjie Lu, and Gaurav Gupta. Deblur-yolo: Real-time object detection with efficient blind motion deblurring. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021.