Prior-based Domain Adaptive Object Detection for Hazy and Rainy Conditions

Vishwanath A. Sindagi*, Poojan Oza*, Rajeev Yasarla, and Vishal M. Patel

Department of Electrical and Computer Engineering,
Johns Hopkins University, 3400 N. Charles St, Baltimore, MD 21218, USA
{vishwanathsindagi,poza2,ryasarl1,vpatel36}@jhu.edu

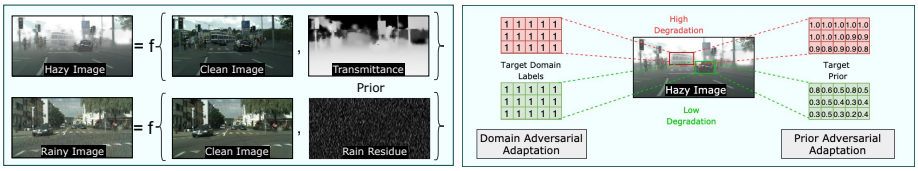
Abstract. Adverse weather conditions such as haze and rain corrupt the quality of captured images, which cause detection networks trained on clean images to perform poorly on these corrupted images. To address this issue, we propose an unsupervised prior-based domain adversarial object detection framework for adapting the detectors to hazy and rainy conditions. In particular, we use weather-specific prior knowledge obtained using the principles of image formation to define a novel prior-adversarial loss. The prior-adversarial loss, which we use to supervise the adaptation process, aims to reduce the weather-specific information in the features, thereby mitigating the effects of weather on the detection performance. Additionally, we introduce a set of residual feature recovery blocks in the object detection pipeline to de-distort the feature space, resulting in further improvements. Evaluations performed on various datasets (Foggy-Cityscapes, Rainy-Cityscapes, RTTS and UFDD) for rainy and hazy conditions demonstrates the effectiveness of the proposed approach.

Keywords: detection, unsupervised domain adaptation, adverse weather, rain, haze

1 Introduction

Object detection [54,12,17,16,31,40,49] is an extensively researched topic in the literature. Despite the success of deep learning based detectors on benchmark datasets [10,9,15,30], they have limited abilities in generalizing to several practical conditions such as adverse weather. This can be attributed mainly to the domain shift in the input images. One approach to solve this issue is to undo the effects of weather conditions by pre-processing the images using existing methods like image dehazing [11,19,61] and/or deraining [28,60,59]. However, these approaches usually involve complex networks and need to be trained separately with pixel-level supervision. Moreover, as noted in [44], these methods additionally involve certain post-processing like gamma correction, which still results in a domain shift, thus prohibiting such approaches from achieving the optimal performance. Like [44], we observed minimal improvements in the detection performance when

* equal contribution



(a) (b)

Fig. 1. (a) Weather conditions can be modeled as function of clean image and the weather-specific prior, which we use to define a novel prior-adversarial loss. (b) Existing adaptation approaches use constant domain label for the entire. Our method uses spatially-varying priors that are directly correlated to the amount of degradations.

we used dehaze/derain methods as a pre-processing step before detection (see Sec. 4). Furthermore, this additional pre-processing results in increased computational overhead at inference, which is not preferable in resource-constrained/real-time applications. Another approach would be to re-train the detectors on datasets that include these adverse conditions. However, creating these datasets often involves high annotation/labeling cost [52].

Recently, a few methods [6,46,42] have attempted to overcome this problem by viewing object detection in adverse weather conditions as an unsupervised domain adaptation task. These approaches consider that the images captured under adverse conditions (target images) suffer from a distribution shift [6,18] as compared to the images on which the detectors are trained (source images). It is assumed that the source images are fully annotated while the target images (with weather-based degradations) are not annotated. They propose different techniques to align the target features with the source features, while training on the source images. These methods are inherently limited in their approach since they employ only the principles of domain adaptation and neglect additional information that is readily available in the case of weather-based degradations.

We consider the following observations about weather-based degradations which have been ignored in the earlier work. ***(i)*** Images captured under weather conditions (such as haze and rain) can be mathematically modeled (see Fig. 1(a), Eq. 8 and 9). For example, a hazy image is modeled by a superposition of a clean image (attenuated by transmission map) and atmospheric light [11,19]. Similarly, a rainy image is modeled as a superposition of a clean image and rain residue [28,59,60] (see Fig. 1(a)). In other words, a weather-affected image contains weather specific information (which we refer to as prior) - transmission map in the case of hazy images and rain residue in the case of rainy images. These weather-specific information/priors cause degradations in the feature space resulting in poor detection performance. Hence, in order to reduce the degradations in the features, it is crucial to make the features weather-invariant by eliminating the weather-specific priors from the features. ***(ii)*** Further, it is important to note that the weather-based degradations are spatially varying and, hence do not affect the features equally at all spatial locations. Since, existing domain-adaptive

detection approaches [6,46,42] label all the locations entirely either as target, they assume that the entire image has undergone constant degradation at all spatial locations (see Fig. 1(b)). This can potentially lead to incorrect alignment, especially in the regions of images where the degradations are minimal.

Motivated by these observations, we define a novel prior-adversarial loss that uses additional knowledge about the target domain (weather-affected images) for aligning the source and target features. Specifically, the proposed loss is used to train a prior estimation network to predict weather-specific prior from the features in the main branch, while simultaneously minimizing the weather-specific information present in the features. This results in weather-invariant features in the main branch, hence, mitigating the effects of weather. Additionally, the proposed use of prior information in the loss function results in spatially varying loss that is directly correlated to the amount of degradation (as shown in Fig. 1(b)). Hence, the use of prior can help avoid incorrect alignment.

Finally, considering that the weather-based degradations cause distortions in the feature space, we introduce a set of residual feature recovery blocks in the object detection pipeline to de-distort the features. These blocks, inspired by residual transfer framework proposed in [20], result in further improvements.

We perform extensive evaluations on different datasets such as Foggy-Cityscapes [44], RTTS [26] and UFDD [36]. Additionally, we create a Rainy-Cityscapes dataset for evaluating the performance different detection methods on rainy conditions. Various experiments demonstrate that the proposed method is able to outperform the existing methods on all the datasets.

2 Related Work

Object detection: Object detection is one of the most researched topics in computer vision. Typical solutions for this problem have evolved from approaches involving sliding window based classification [54,8] to the latest anchor-based convolutional neural network approaches [40,39,31]. Ren *et al.*[40] pioneered the popular two stage Faster-RCNN approach. Several works have proposed single stage frameworks such as SSD [31], YOLO [39] *etc*, that directly predict the object labels and bounding box co-ordinates. Following the previous work [6,46,42,24,23], we use Faster-RCNN as our base model.

Domain-adaptive object detection in adverse conditions: Compared to the problem of general detection, detection in adverse weather conditions is relatively less explored. Existing methods, [6,46,42,24] have attempted to address this task from a domain adaptation perspective. Chen *et al.*[6] assumed that the adversarial weather conditions result in domain shift, and they overcome this by proposing a domain adaptive Faster-RCNN approach that tackles domain shift on image-level and instance-level. Following the similar argument of domain shift, Shan *et al.*[46] proposed to perform joint adaptation at image level using the Cycle-GAN framework [64] and at feature level using conventional domain adaptation losses. Saito *et al.*[42] proposed to perform strong alignment of the local features and weak alignment of the global features. Kim *et al.*[24]

diversified the labeled data, followed by adversarial learning with the help of multi-domain discriminators. Cai *et al.*[5] addressed this problem in the semi-supervised setting using mean teacher framework. Zhu *et al.*[65] proposed region mining and region-level alignment in order to correctly align the source and target features. Roychowdhury *et al.*[41] adapted detectors to a new domain assuming availability of large number of video data from the target domain. These video data are used to generate pseudo-labels for the target set, which are further employed to train the network. Most recently, Khodabandeh *et al.*[23] formulated the domain adaptation training with noisy labels. Specifically, the model is trained on the target domain using a set of noisy bounding boxes that are obtained by a detection model trained only in the source domain.

3 Proposed Method

We assume that labeled clean data ($\{x_i^s, y_i^s\}_{i=1}^{n_s}$) from the source domain ($\mathcal{S}$) and unlabeled weather-affected data from the target domain ($\mathcal{T}$) are available. Here, y_i^s refers to all bounding box annotations and respective category label for the corresponding clean image x_i^s, x_i^t refers to the weather-affected image, n_s is the total number of samples in the source domain ($\mathcal{S}$) and n_t is the total number of samples in the target domain ($\mathcal{T}$). Our goal is to utilize the available information in both source and target domains to learn a network that lessens the effect of weather-based conditions on the detector. The proposed method contains three network modules – detection network, prior estimation network (PEN) and residual feature recovery block (RFRB). Fig. 2 gives an overview of the proposed model. During source training, a source image (clean image) is passed to the detection network and the weights are learned by minimizing the detection loss, as shown in Fig. 2 with the source pipeline. For target training, a target image (weather-affected image) is forwarded through the network as shown in Fig. 2 by the target pipeline. As discussed earlier, weather-based degradations cause distortions in the feature space for the target images. In an attempt to de-distort these features, we introduce a set of residual feature recovery blocks in the target pipeline as shown in Fig. 2. This model is inspired from residual transfer framework proposed in [33] and is used to model residual features. The proposed PEN aids the detection network in adapting to the target domain by providing feedback through adversarial training using the proposed prior adversarial loss. In the following subsections, we briefly review the backbone network, followed by a discussion on the proposed prior-adversarial loss and residual feature recovery blocks.

3.1 Detection Network

Following the existing domain adaptive detection approaches [6,46,42], we base our method on the Faster-RCNN [40] framework. Faster-RCNN is among the first end-to-end CNN-based object detection methods and uses anchor-based strategy to perform detection and classification. For this paper we decompose the

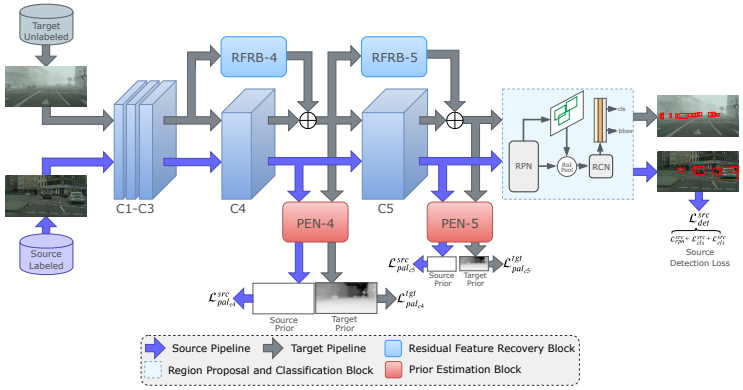


Fig. 2. Overview of the proposed adaptation method. We use prior adversarial loss to supervise the domain discriminators. For the source pipeline, additional supervision is provided by detection loss. For target pipeline, feed-forward through the detection network is modified by the residual feature recovery blocks.

Faster-RCNN network into three network modules: feature extractor network ($\mathcal{F}$), region proposal network (RPN) stage and region classification network (RCN). The arrangement of these modules are shown in the Fig. 2 with VGG model architecture as base network. Here, the feature extractor network consists of first five conv blocks of VGG and region classification network module is composed of fully connected layers of VGG. The region proposal network uses output of feature extractor network to generate a set of candidate object regions in a class agnostic way. Features corresponding to these candidates are pooled from the feature extractor and are forwarded through the region classification network to get the object classifications and bounding box refinements. Since we have access to the source domain images and their corresponding ground truth, these networks are trained to perform detection on the source domain by minimizing the following loss function,

$$\min_{\mathcal{F},\,\mathcal{G}} \quad \mathcal{L}_{det}^{src}\,, \quad \text{where} \tag{1}$$

$$\mathcal{L}_{det}^{src} \;=\; \mathcal{L}_{rpn}^{src} + \mathcal{L}_{bbox}^{src} + \mathcal{L}_{rcn}^{src}. \tag{2}$$

Here, $\mathcal{G}$ represents both region proposal and region classification networks, $\mathcal{L}_{rpn}^{src}$ denotes the region proposal loss, $\mathcal{L}_{bbox}^{src}$ denotes the bounding-box regression loss and $\mathcal{L}_{rcn}^{src}$ denotes the region classification loss. The details of these individual loss components can be found in [40].

3.2 Prior-adversarial Training

As discussed earlier, weather-affected images, contain domain specific information. These images typically follow mathematical models of image degradation (see Fig.

1(a), Eq. 8 and Eq. 9). We refer to this domain specific information as a *prior*. Detailed discussion about prior for haze and rain is provided later in the section. We aim to exploit these priors about the weather domain to better adapt the detector for weather affected images. To achieve that, we propose a prior-based adversarial training approach using prior estimation network (PEN) and prior adversarial loss (PAL).

Let $\mathcal{P}_l$ be PEN module introduced after the l^{th} conv block of $\mathcal{F}$ and let Z_{il}^{src} be the corresponding domain specific prior for any image, $x_i^s \in \mathcal{S}$. Then the PAL for the source domain is defined as follows,

$$\mathcal{L}_{pal_{cl}}^{src} = \frac{1}{n_s UV} \sum_{i=1}^{n_s} \sum_{j=1}^{U} \sum_{k=1}^{V} (Z_{il}^{src} - \mathcal{P}_l(\mathcal{F}_l(x_i^s)))_{jk}^2, \tag{3}$$

where, U and V are height and width of domain specific prior Z_{il}^{src} and output feature $\mathcal{F}_l(x_i^s)$. Z_{il}^{src} denotes the source image prior, scaled down from image-level prior to match the scale at l^{th} conv block. Similarly, PAL for the target domain images, $x_i^t \in \mathcal{T}$, with the corresponding prior Z_{il}^{tgt} can be defined as,

$$\mathcal{L}_{pal_{cl}}^{tgt} = \frac{1}{n_t UV} \sum_{i=1}^{n_t} \sum_{j=1}^{U} \sum_{k=1}^{V} (Z_{il}^{tgt} - \mathcal{P}_l(\mathcal{F}_l(x_i^t)))_{jk}^2, \tag{4}$$

where, we apply PAL after conv4 (l=4) and conv5 (l=5) block (as shown in Fig. 2). Hence, the final source and target adversarial losses can be given as,

$$\mathcal{L}_{pal}^{src} = \frac{1}{2}(\mathcal{L}_{pal_{c5}}^{src} + \mathcal{L}_{pal_{c4}}^{src}), \tag{5}$$

$$\mathcal{L}_{pal}^{tgt} = \frac{1}{2}(\mathcal{L}_{pal_{c5}}^{tgt} + \mathcal{L}_{pal_{c4}}^{tgt}). \tag{6}$$

The prior estimation networks ($\mathcal{P}_5$ and $\mathcal{P}_4$) predict the weather-specific prior from the features extracted from $\mathcal{F}$. However, the feature extractor network $\mathcal{F}$ is trained to fool the PEN modules by producing features that are weather-invariant (free from weather-specific priors) and prevents the PEN modules from correctly estimating the weather-specific prior. Since, this type of training includes prior prediction and is also reminiscent of the adversarial learning used in domain adaptation, we term this loss as prior-adversarial loss. At convergence, the feature extractor network $\mathcal{F}$ should have devoid itself from any weather-specific information and as a result both prior estimation networks $\mathcal{P}_5$ and $\mathcal{P}_4$ should not be able to correctly estimate the prior. *Note that our goal at convergence is not to estimate the correct prior, but rather to learn weather-invariant features so that the detection network is able to generalize well to the target domain.* This training procedure can be expressed as the following optimization,

$$\max_{\mathcal{F}} \min_{\mathcal{P}} \quad \mathcal{L}_{pal}^{src} + \mathcal{L}_{pal}^{tgt}. \tag{7}$$

Furthermore, in the conventional domain adaptation, a single label is assigned for entire target image to train the domain discriminator (Fig. 1)(c)). By doing this, it is assumed that the entire image has undergone a constant domain shift. However this is not true in the case of weather-affected images, where degradations vary spatially (Fig. 1)(b)). In such cases, the assumption of constant domain shift leads to incorrect alignment especially in the regions of minimal degradations. Incorporating the weather-specific priors overcomes this issue as these priors are spatially varying and are directly correlated with the amount of degradations. Hence, utilizing the weather-specific prior results in better alignment.

Haze prior The effect of haze on images has been extensively studied in the literature [11,19,61,29,4,62,63]. Most existing image dehazing methods rely on the atmospheric scattering model for representing image degradations under hazy conditions and is defined as,

$$I(z) = J(z)t(z) + A(z)(1 - t(z)), \tag{8}$$

where I is the observed hazy image, J is the true scene radiance, A is the global atmospheric light, indicating the intensity of the ambient light, t is the transmission map and z is the pixel location. The transmission map is a distance-dependent factor that affects the fraction of light that reaches the camera sensor. When the atmospheric light A is homogeneous, the transmission map can be expressed as $t(z) = e^{-\beta d(z)}$, where β represents the attenuation coefficient of the atmosphere and d is the scene depth.

Typically, existing dehazing methods first estimate the transmission map and the atmospheric light, which are then used in Eq. (8) to recover the observed radiance or clean image. The transmission map contains important information about the haze domain, specifically representing the light attenuation factor. We use this transmission as a domain prior for supervising the prior estimation (PEN) while adapting to hazy conditions. *Note that no additional human annotation efforts are required for obtaining the haze prior.*

Rain prior Similar to dehazing, image deraining methods [28,59,60,27,58] also assume a mathematical model to represent the degradation process and is defined as follows,

$$I(z) = J(z) + R(z), \tag{9}$$

where I is the observed rainy image, J is the desired clean image, and R is the rain residue. This formulation models rainy image as a superposition of the clean background image with the rain residue. The rain residue contains domain specific information about the rain for a particular image and hence, can be used as a domain specific prior for supervising the prior estimation network (PEN) while adapting to rainy conditions. *Similar to the haze, we avoid the use of expensive human annotation efforts for obtaining the rain prior.*

In both cases discussed above (haze prior and rain prior), we do not use any ground-truth labels to estimate respective priors. Hence, our overall approach still falls into the category of unsupervised adaptation. Furthermore, these priors can be pre-computed for the training images to reduce the computational overhead during the learning process. Additionally, the prior computation is not required during inference and hence, the proposed adaptation method does not result in any computational overhead.

3.3 Residual Feature Recovery Block

As discussed earlier, weather-degradations introduce distortions in the feature space. In order to aid the de-distortion process, we introduce a set of residual feature recovery blocks (RFRBs) in the target feed-forward pipeline. This is inspired from the residual transfer network method proposed in [33]. Let $\Delta\mathcal{F}_l$ be the residual feature recovery block at the l^{th} conv block. The target domain image feedforward is modified to include the residual feature recovery block. For $\Delta\mathcal{F}_l$ the feed-forward equation at the l^{th} conv block can be written as,

$$\hat{\mathcal{F}}_l(x_i^t) \;=\; \mathcal{F}_l(x_i^t) \;+\; \Delta\mathcal{F}_l(\mathcal{F}_{l-1}(x_i^t)), \tag{10}$$

where, $\mathcal{F}_l(x_i^t)$ indicates the feature extracted from the l^{th} conv block for any image x_i^t sampled from the target domain using the feature extractor network $\mathcal{F}$, $\Delta\mathcal{F}_l(\mathcal{F}_{l-1}(x_i^t))$ indicates the residual features extracted from the output $l-1^{th}$ conv block, and $\hat{\mathcal{F}}_l(x_i^t)$ indicates the feature extracted from the l^{th} conv block for any image $x_i^t \in \mathcal{T}$ with RFRB modified feedforward. The RFRB modules are also illustrated in Fig. 2, as shown in the target feedforward pipeline. It has no effect on source feedforward pipeline. In our case, we utilize RFRB at both conv4 ($\Delta\mathcal{F}_4$) and conv5 ($\Delta\mathcal{F}_5$) blocks. Additionally, the effect of residual feature is regularized by enforcing the norm constraints on the residual features. The regularization loss for RFRBs, $\Delta\mathcal{F}_4$ and $\Delta\mathcal{F}_5$ is defined as,

$$\mathcal{L}_{reg} = \frac{1}{n_t}\sum_{i=1}^{n_t}\sum_{l=4,5} \|\Delta\mathcal{F}_l(\mathcal{F}_{l-1}(x_i^t))\|_1, \tag{11}$$

3.4 Overall Loss

The overall loss for training the network is defined as,

$$\max_{\mathcal{P}} \min_{\mathcal{F},\Delta\mathcal{F},\mathcal{G}} \quad \mathcal{L}_{det}^{src} - \mathcal{L}_{adv} + \lambda\mathcal{L}_{reg}, \quad \text{where} \tag{12}$$

$$\mathcal{L}_{adv} = \frac{1}{2}(\mathcal{L}_{pal}^{src} + \mathcal{L}_{pal}^{tgt}). \tag{13}$$

Here, $\mathcal{F}$ represents the feature extractor network, $\mathcal{P}$ denotes both prior estimation network employed after conv4 and conv5 blocks, i.e., $\mathcal{P}=\{\mathcal{P}_5, \mathcal{P}_4\}$, and $\Delta\mathcal{F}=\{\Delta\mathcal{F}_4, \Delta\mathcal{F}_5\}$ represents RFRB at both conv4 and conv5 blocks. Also, $\mathcal{L}_{det}^{src}$ is the source detection loss, $\mathcal{L}_{reg}$ is the regularization loss, and $\mathcal{L}_{adv}$ is the overall adversarial loss used for prior-based adversarial training.

4 Experiments and Results

4.1 Implementation details

We follow the training protocol of [42,6] for training the Faster-RCNN network. The backbone network for all experiments is VGG16 network [48]. We model the residuals using RFRB for the convolution blocks C4 and C5 of the VGG16 network. The PA loss is applied to only these conv blocks modeled with RFRBs. The PA loss is designed based on the adaptation setting (Haze or Rain). The parameters of the first two conv blocks are frozen similar to [42,6]. The detailed network architecture for RFRBs, PEN and the discriminator are provided in supplementary material. During training, we set shorter side of the image to 600 with ROI alignment. We train all networks for 70K iterations. For the first 50K iterations, the learning rate is set equal to 0.001 and for the last 20K iterations it is set equal to 0.0001. We report the performance based on the trained model after 70K iterations. We set λ equal to 0.1 for all experiments.

In addition to comparison with recent methods, we also perform an ablation study where we evaluate the following configurations to analyze the effectiveness of different components in the network. Note that we progressively add additional components which enables us to gauge the performance improvements obtained by each of them,

- **FRCNN:** Source only baseline experiment where Faster-RCNN is trained on the source dataset.
- **FRCNN+$\mathbf{D_5}$:** Domain adaptation baseline experiment consisting of Faster-RCNN with domain discriminator after conv5 supervised by the domain adversarial loss.
- **FRCNN+$\mathbf{D_5}$+$\mathbf{R_5}$:** Starting with FRCNN+D_5 as the base configuration, we add an RFRB block after conv4 in the Faster-RCNN. This experiment enables us to understand the contribution of the RFRB block.
- **FRCNN+$\mathbf{P_5}$+$\mathbf{R_5}$:** We start with FRCNN+D_5+R_5 configuration and replace domain discriminator and domain adversarial loss with prior estimation network (PEN) and prior adversarial loss (PAL). With this experiment, we show the importance of training with the proposed prior-adversarial loss.
- **FRCNN+$\mathbf{P_{45}}$+$\mathbf{R_{45}}$:** Finally, we perform the prior-based feature alignment at two scales: conv4 and conv5. Starting with FRCNN+P_5+R_5 configuration, we add an RFRB block after conv3 and a PEN module after conv4. This experiment corresponds to the configuration depicted in Fig. 2. This experiment demonstrates the efficacy of the overall method in addition to establishing the importance of aligning features at multiple levels in the network.

Following the protocol set by the existing methods [6,46,42], we use mean average precision (mAP) scores for performance comparison.

4.2 Adaptation to hazy conditions

In this section, we present the results corresponding to adaptation to hazy conditions on the following datasets: (i) Cityscapes $\rightarrow$ Foggy-Cityscapes [44], (ii)

(a) (b)

Fig. 3. Detection results on Foggy-Cityscapes. (a) DA-Faster RCNN [6]. (b) Proposed method. The bounding boxes are colored based on the detector confidence. DA-Faster-RCNN produces detections with low confidence in addition to missing the truck class. Our method is able to output high confidence detections without missing any objects.

Cityscapes → RTTS [25], and (iii) WIDER [56] → UFDD-Haze [36]. In the first two experiments, we consider Cityscapes [7] as the source domain. Note that the Cityscapes dataset contains images captured in clear weather conditions.

Cityscapes → Foggy-Cityscapes: In this experiment, we adapt from Cityscapes to Foggy-Cityscapes [44]. The Foggy-Cityscapes dataset was recently proposed in [44] to study the detection algorithms in the case of hazy weather conditions. Foggy-Cityscapes is derived from Cityscapes dataset by simulating fog on the clear weather images of Cityscapes. Both Cityscapes and Foggy-Cityscapes have the same number of categories which include, car, truck, motorcycle/bike, train, bus, rider and person. Similar to [6], [42], we utilize 2975 images of both Cityscapes and Foggy-Cityscapes for training. Note that we use annotations only from the source dataset (Cityscapes) for training the detection pipeline. For evaluation we consider a non overlapping validation set of 500 images provided by the Foggy-Cityscapes dataset.

We compare the proposed method with two categories of approaches: *(i) Dehaze+Detect:* Here, we employ dehazing network as pre-processing step and perform detection using Faster-RCNN trained on source (clean) images. For pre-processing, we chose two recent dehazing algorithms: DCPDN [61] and Grid-Dehaze [32]. *(i) DA-based methods:* Here, we compare with following recent domain-adaptive detection approaches: DA-Faster [6], SWDA [42], DiversifyMatch [24], Mean Teacher with Object Relations (MTOR) [5], Selective Cross-Domain Alignment (SCDA) [65] and Noisy Labeling [23]. The corresponding results are presented in Table 1.

It can be observed from Table 1, that the performance of source-only training of Faster-RCNN is in general poor in the hazy conditions. Adding DCPDN and Gird-Dehaze as preprocessing step improves the performance by ∼2% and ∼4%, respectively. Compared to the domain-adaptive detection approaches, pre-

Table 1. Performance comparison for the Cityscapes → Foggy-Cityscapes experiment.

Method		prsn	rider	car	truc	bus	train	bike	bcycle	mAP
Baseline	FRCNN [40]	25.8	33.7	35.2	13.0	28.2	9.1	18.7	31.4	24.4
Dehaze	DCPDN [61]	27.9	36.2	35.2	16.0	28.3	10.2	24.6	32.5	26.4
	Grid-Dehaze [32]	29.7	40.4	40.3	21.3	30.0	9.1	25.6	36.7	29.2
DA-Methods	DAFaster [6]	25.0	31.0	40.5	22.1	35.3	20.2	20.0	27.1	27.6
	SCDA [65]	33.5	38.0	48.5	26.5	39.0	23.3	28.0	33.6	33.8
	SWDA [42]	29.9	42.3	43.5	24.5	36.2	32.6	30.0	35.3	34.3
	DM [24]	30.8	40.5	44.3	**27.2**	38.4	34.5	28.4	32.2	34.6
	MTOR [5]	30.6	41.4	44.0	21.9	38.6	**40.6**	28.3	35.6	35.1
	NL [23]	35.1	42.1	49.2	30.1	45.3	26.9	26.8	36.0	36.5
Ours	FRCNN+D_5	30.9	38.5	44.0	19.6	32.9	17.9	24.1	32.4	30.0
	FRCNN+D_5+R_5	32.8	44.7	49.9	22.3	31.7	17.3	26.9	37.5	32.9
	FRCNN+P_5+R_5	33.4	42.8	50.0	24.2	40.8	30.4	33.1	37.5	36.5
	FRCNN+P_{45}+R_{45}	**36.4**	**47.3**	**51.7**	22.8	**47.6**	34.1	**36.0**	**38.7**	**39.3**

processing + detection results in lower performance gains. This is because even after applying dehazing there still remains some domain shift as discussed in Sec. 1. Hence, using adaptation would be a better approach for mitigating the domain shift. Here, the use of simple domain adaptation [14] (FRCNN+D_5) improves the source-only performance. The addition of $RFRB_5$ (FRCNN+D_5+R_5) results in further improvements, thus indicating the importance of RFRB blocks. However, the conventional domain adaptation loss assumes constant domain shift across the entire image, resulting in incorrect alignment. The use of prior-adversarial loss (FRCNN+P_5+R_5) overcomes this issue. We achieved 3.6% improvement in overall mAP scores, thus demonstrating the effectiveness of the proposed prior-adversarial training. Note that, FRCNN+P_5+R_5 baseline achieves comparable performance with state-of-the-art. Finally, by performing prior-adversarial adaptation at an additional scale (FRCNN+P_{45}+R_{45}), we achieve further improvements which surpasses the existing best approach [23] by 2.8%. Fig. 3 shows sample qualitative detection results corresponding to the images from Foggy-Cityscapes. Results for the proposed method are compared with DA-Faster-RCNN [6]. It can be observed that the proposed method is able to generate comparatively high quality detections.

We summarize our observations as follows: (i) Using dehazing as a pre-processing step results in minimal improvements over the baseline Faster-RCNN. Domain adaptive approaches perform better in general. (ii) The proposed method outperforms other methods in the overall scores while achieving the best performance in most of the classes. See supplementary material for more ablations.

Cityscapes → RTTS: In this experiment, we adapt from Cityscapes to the RTTS dataset [25] . RTTS is a subset of a larger RESIDE dataset [25], and it contains 4,807 unannotated and 4,322 annotated real-world hazy images covering mostly traffic and driving scenarios. We use the unannotated 4,807 images for training the domain adaptation process. The evaluation is performed on the annotated 4,322 images. RTTS has total five categories, namely motorcycle/bike,

Table 2. Performance comparison for the Cityscapes → RTTS experiment.

Method		prsn	car	bus	bike	bcycle	mAP
Baseline	FRCNN [40]	46.6	39.8	11.7	19.0	37.0	30.9
Dehaze	DCPDN [61]	**48.7**	39.5	12.9	19.7	37.5	31.6
	Grid-Dehaze [32]	29.7	25.4	10.9	13.0	21.4	20.0
DA	DAFaster [6]	37.7	48.0	14.0	**27.9**	36.0	32.8
	SWDA [42]	42.0	46.9	15.8	25.3	37.8	33.5
Ours	Proposed	37.4	**54.7**	**17.2**	22.5	**38.5**	**34.1**

Table 3. Results (mAP) of the adaptation experiments from WIDER-Face to UFDD Haze and Rain.

Method	UFDD-Haze	UFDD-Rain
FRCNN [40]	46.4	54.8
DAFaster [6]	52.1	58.2
SWDA [42]	55.5	60.0
Proposed	**58.5**	**62.1**

person, bicycle, bus and car. This dataset is the largest available dataset for object detection under real world hazy conditions.

In Table 2, the results of the proposed method are compared with Faster-RCNN [40], DA-Faster [6] and SWDA [42] and the dehaze+detection baseline as well. For RTTS dataset, the pre-processing with DCPDN improves the Faster-RCNN performance by ∼1%. Surprisingly, Grid-Dehaze does not help the Faster-RCNN baseline and results in even worse performance. Whereas, the proposed method achieves an improvement of 3.1% over the baseline Faster-RCNN (source-only training), while outperforming the other recent methods.

WIDER-Face → UFDD-Haze: Recently, Nada *et al.*[36] published a benchmark face detection dataset which consists of real-world images captured under different weather-based conditions such as haze and rain. Specifically, this dataset consists of 442 images under the haze category. Since, face detection is closely related to the task of object detection, we evaluate our framework by adapting from WIDER-Face [56] dataset to UFDD-Haze dataset. WIDER-Face is a large-scale face detection dataset with approximately 32,000 images and 199K face annotations. The results corresponding to this adaptation experiment are shown in Table 3. It can be observed from this table that the proposed method achieves better performance as compared to the other methods.

4.3 Adaptation to rainy conditions

In this section, we present the results of adaptation to rainy conditions. Due to lack of appropriate datasets for this particular setting, we create a new rainy dataset called Rainy-Cityscapes and it is derived from Cityscapes. It has the same number of images for training and validation as Foggy-Cityscapes. First,

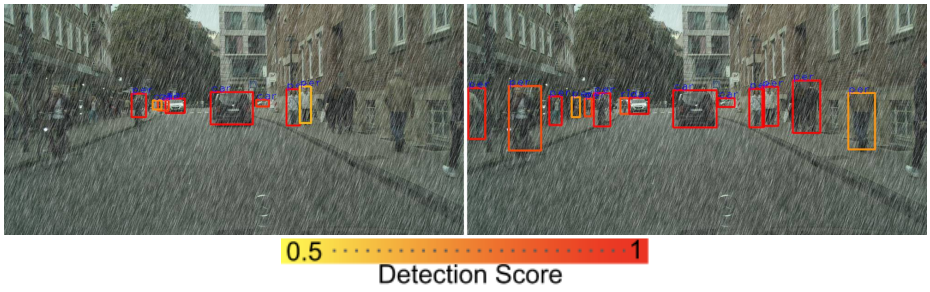
Table 4. Performance comparison for the Cityscapes → Rainy-Cityscapes experiment.

Method		prsn	rider	car	truc	bus	train	bike	bcycle	mAP
Baseline	FRCNN	21.6	19.5	38.0	12.6	30.1	24.1	12.9	15.4	21.8
Derain	DDN [13]	27.1	30.3	50.7	23.1	39.4	18.5	21.2	24.0	29.3
	SPANet [55]	24.9	28.9	48.1	21.4	34.8	16.8	17.6	20.8	26.7
DA	DAFaster [6]	26.9	28.1	50.6	23.2	39.3	4.7	17.1	20.2	26.3
	SWDA [42]	29.6	**38.0**	52.1	27.9	**49.8**	28.7	24.1	25.4	34.5
Ours	FRCNN+D_5	29.1	34.8	52.0	22.0	41.8	20.4	18.1	23.3	30.2
	FRCNN+D_5+R_5	28.8	33.1	51.7	22.3	41.8	24.9	22.2	24.6	31.2
	FRCNN+P_5+R_5	29.7	34.3	52.5	23.6	47.9	32.5	24.0	25.5	33.8
	FRCNN+P_{45}+R_{45}	**31.3**	34.8	**57.8**	**29.3**	48.6	**34.4**	**25.4**	**27.3**	**36.1**

we discuss the simulation process used to create the dataset, followed by a discussion of the evaluation and comparison of the proposed method with other methods.

Rainy-Cityscapes: Similar to Foggy-Cityscapes, we use a subset of 3475 images from Cityscapes to create synthetic rain dataset. Using [1], several masks containing artificial rain streaks are synthesized. The rain streaks are created using different Gaussian noise levels and multiple rotation angles between 70° and 110°. Next, for every image in the subset of the Cityscapes dataset, we pick a random rain mask and blend it onto the image to generate the synthetic rainy image. More details and example images are provided in supplementary material.

Cityscapes → Rainy-Cityscapes: In this experiment, we adapt from Cityscapes to Rainy-Cityscapes. We compare the proposed method with recent methods such as DA-Faster [6] and SWDA [42]. Additionally, we also evaluate performance of two derain+detect baselines, where state of the art methods such as DDN [13] and SPANet [55] are used as a pre-processing step to the Faster-RCNN trained on source (clean) images. From the Table 4 we observe that such methods provide reasonable improvements over the Faster-RCNN baseline. However, the performance gains are much lesser as compared to adaptation methods, for the reasons discussed in the earlier sections (Sec. 1, Sec. 4.2). Also, it can be observed from Table 4, that the proposed method outperforms the other methods by a significant margin. Additionally, we present the results of the ablation study consisting of the experiments listed in Sec. 4.1. The introduction of domain adaptation loss significantly improves the source only Faster-RCNN baseline, resulting in approximately 9% improvement for FRCNN+D_5 baseline in Table 4. This performance is further improved by 1% with the help of residual feature recovery blocks as shown in FRCNN+D_5+R_5 baseline. When domain adversarial training is replaced with prior adversarial training with PAL, i.e. FRCNN+P_5+R_5 baseline, we observe 2.5% improvements, showing effectiveness of the proposed training methodology. Finally, by performing prior adversarial training at multiple scales, the proposed method FRCNN+P_{45}+R_{45} observes approximately 2% improvements and also outperforms the next best method SWDA [42] by 1.6%. Fig. 4 illustrates sample detection results obtained using the proposed method as compared



(a) (b)

Fig. 4. Detection results on Rainy-Cityscapes. (a) DA-Faster RCNN [6]. (b) Proposed method. The bounding boxes are colored based on the detector confidence. DA-Faster-RCNN misses several objects. Our method is able to output high confidence detections without missing any objects.

to a recent method [6]. The proposed method achieves superior quality detections.

WIDER-Face → UFDD-Rain: In this experiment, we adapt from WIDER-Face to UFDD-Rain [36]. The UFDD-Rain dataset consists of 628 images collected under rainy conditions. The results of the proposed method as compared to the other methods are shown in Table 3. It can be observed that the proposed method outperforms the source only training by 7.3%.We provide additional details about the proposed method including results and analysis in the supplementary material.

5 Conclusions

We addressed the problem of adapting object detectors to hazy and rainy conditions. Based on the observation that these weather conditions cause degradations that can be mathematically modeled and cause spatially varying distortions in the feature space, we propose a novel prior-adversarial loss that aims at producing weather-invariant features. Additionally, a set of residual feature recovery blocks are introduced to learn residual features that can aid efficiently aid the adaptation process. The proposed framework is evaluated on several benchmark datasets such as Foggy-Cityscapes, RTTS and UFDD. Through extensive experiments, we showed that our method achieves significant gains over the recent methods in all the datasets.

Acknowledgement

This work was supported by the NSF grant 1910141.

References

1. https://www.photoshopessentials.com/photo-effects/photoshop-weather-effects-rain/
2. Abavisani, M., Patel, V.M.: Domain adaptive subspace clustering. In: 27th British Machine Vision Conference, BMVC 2016 (2016)
3. Abavisani, M., Patel, V.M.: Adversarial domain adaptive subspace clustering. In: 2018 IEEE 4th International Conference on Identity, Security, and Behavior Analysis (ISBA). pp. 1–8. IEEE (2018)
4. Ancuti, C., Ancuti, C.O., Timofte, R.: Ntire 2018 challenge on image dehazing: Methods and results. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 891–901 (2018)
5. Cai, Q., Pan, Y., Ngo, C.W., Tian, X., Duan, L., Yao, T.: Exploring object relation in mean teacher for cross-domain detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 11457–11466 (2019)
6. Chen, Y., Li, W., Sakaridis, C., Dai, D., Gool, L.V.: Domain adaptive faster r-cnn for object detection in the wild. 2018 IEEE Conference on Computer Vision and Pattern Recognition pp. 3339–3348 (2018)
7. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016)
8. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: international Conference on computer vision & Pattern Recognition (CVPR'05). vol. 1, pp. 886–893. IEEE Computer Society (2005)
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
10. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**(2), 303–338 (2010)
11. Fattal, R.: Single image dehazing. ACM transactions on graphics (TOG) **27**(3), 72 (2008)
12. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE transactions on pattern analysis and machine intelligence **32**(9), 1627–1645 (2010)
13. Fu, X., Huang, J., Zeng, D., Huang, Y., Ding, X., Paisley, J.: Removing rain from single images via a deep detail network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3855–3863 (2017)
14. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. arXiv preprint arXiv:1409.7495 (2014)
15. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. The International Journal of Robotics Research **32**(11), 1231–1237 (2013)
16. Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 1440–1448 (2015)
17. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 580–587 (2014)
18. Gopalan, R., Li, R., Chellappa, R.: Domain adaptation for object recognition: An unsupervised approach. In: 2011 international conference on computer vision. pp. 999–1006. IEEE (2011)

19. He, K., Sun, J., Tang, X.: Single image haze removal using dark channel prior. IEEE transactions on pattern analysis and machine intelligence **33**(12), 2341–2353 (2011)
20. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
21. Hoffman, J., Tzeng, E., Park, T., Zhu, J.Y., Isola, P., Saenko, K., Efros, A.A., Darrell, T.: Cycada: Cycle-consistent adversarial domain adaptation. arXiv preprint arXiv:1711.03213 (2017)
22. Hu, L., Kan, M., Shan, S., Chen, X.: Duplex generative adversarial network for unsupervised domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1498–1507 (2018)
23. Khodabandeh, M., Vahdat, A., Ranjbar, M., Macready, W.G.: A robust learning approach to domain adaptive object detection. arXiv preprint arXiv:1904.02361 (2019)
24. Kim, T., Jeong, M., Kim, S., Choi, S., Kim, C.: Diversify and match: A domain adaptive representation learning paradigm for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 12456–12465 (2019)
25. Li, B., Ren, W., Fu, D., Tao, D., Feng, D., Zeng, W., Wang, Z.: Benchmarking single-image dehazing and beyond. IEEE Transactions on Image Processing **28**(1), 492–505 (2019)
26. Li, B., Ren, W., Fu, D., Tao, D., Feng, D.D., Zeng, W., Wang, Z.: Benchmarking single-image dehazing and beyond. IEEE Transactions on Image Processing **28**, 492–505 (2018)
27. Li, S., Ren, W., Zhang, J., Yu, J., Guo, X.: Single image rain removal via a deep decomposition–composition network. Computer Vision and Image Understanding (2019)
28. Li, Y., Tan, R.T., Guo, X., Lu, J., Brown, M.S.: Rain streak removal using layer priors. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 2736–2744 (2016)
29. Li, Y., You, S., Brown, M.S., Tan, R.T.: Haze visibility enhancement: A survey and quantitative benchmarking. Computer Vision and Image Understanding **165**, 1–16 (2017)
30. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
31. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: European conference on computer vision. pp. 21–37. Springer (2016)
32. Liu, X., Ma, Y., Shi, Z., Chen, J.: Griddehazenet: Attention-based multi-scale network for image dehazing. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 7314–7323 (2019)
33. Long, M., Zhu, H., Wang, J., Jordan, M.I.: Unsupervised domain adaptation with residual transfer networks. In: Advances in Neural Information Processing Systems. pp. 136–144 (2016)
34. Long, M., Zhu, H., Wang, J., Jordan, M.I.: Deep transfer learning with joint adaptation networks. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. pp. 2208–2217. JMLR. org (2017)

35. Murez, Z., Kolouri, S., Kriegman, D., Ramamoorthi, R., Kim, K.: Image to image translation for domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4500–4509 (2018)
36. Nada, H., Sindagi, V.A., Zhang, H., Patel, V.M.: Pushing the limits of unconstrained face detection: a challenge dataset and baseline results. arXiv preprint arXiv:1804.10275 (2018)
37. Patel, V.M., Gopalan, R., Li, R., Chellappa, R.: Visual domain adaptation: A survey of recent advances. IEEE signal processing magazine **32**(3), 53–69 (2015)
38. Perera, P., Abavisani, M., Patel, V.M.: In2i: Unsupervised multi-image-to-image translation using generative adversarial networks. In: 2018 24th International Conference on Pattern Recognition (ICPR). pp. 140–146. IEEE (2018)
39. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779–788 (2016)
40. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems. pp. 91–99 (2015)
41. RoyChowdhury, A., Chakrabarty, P., Singh, A., Jin, S., Jiang, H., Cao, L., Learned-Miller, E.: Automatic adaptation of object detectors to new domains using self-training. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 780–790 (2019)
42. Saito, K., Ushiku, Y., Harada, T., Saenko, K.: Strong-weak distribution alignment for adaptive object detection. CoRR **abs/1812.04798** (2018)
43. Saito, K., Watanabe, K., Ushiku, Y., Harada, T.: Maximum classifier discrepancy for unsupervised domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3723–3732 (2018)
44. Sakaridis, C., Dai, D., Gool, L.V.: Semantic foggy scene understanding with synthetic data. International Journal of Computer Vision **126**, 973–992 (2018)
45. Sankaranarayanan, S., Balaji, Y., Castillo, C.D., Chellappa, R.: Generate to adapt: Aligning domains using generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8503–8512 (2018)
46. Shan, Y., Feng Lu, W., Meng Chew, C.: Pixel and feature level based domain adaption for object detection in autonomous driving (09 2018)
47. Shu, R., Bui, H.H., Narui, H., Ermon, S.: A dirt-t approach to unsupervised domain adaptation. arXiv preprint arXiv:1802.08735 (2018)
48. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
49. Sindagi, V., Patel, V.: Dafe-fd: Density aware feature enrichment for face detection. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 2185–2195. IEEE (2019)
50. Sindagi, V.A., Srivastava, S.: Domain adaptation for automatic oled panel defect detection using adaptive support vector data description. International Journal of Computer Vision **122**(2), 193–211 (2017)
51. Sindagi, V.A., Yasarla, R., Babu, D.S., Babu, R.V., Patel, V.M.: Learning to count in the crowd from limited labeled data. arXiv preprint arXiv:2007.03195 (2020)
52. Sindagi, V.A., Yasarla, R., Patel, V.M.: Pushing the frontiers of unconstrained crowd counting: New dataset and benchmark method. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1221–1231 (2019)

53. Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7167–7176 (2017)
54. Viola, P., Jones, M., et al.: Rapid object detection using a boosted cascade of simple features. CVPR (1) **1**, 511–518 (2001)
55. 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 Conference on Computer Vision and Pattern Recognition. pp. 12270–12279 (2019)
56. Yang, S., Luo, P., Loy, C.C., Tang, X.: Wider face: A face detection benchmark. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5525–5533 (2016)
57. 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 (CVPR) (June 2020)
58. You, S., Tan, R.T., Kawakami, R., Mukaigawa, Y., Ikeuchi, K.: Adherent raindrop modeling, detectionand removal in video. IEEE transactions on pattern analysis and machine intelligence **38**(9), 1721–1733 (2015)
59. Zhang, H., Patel, V.M.: Density-aware single image de-raining using a multi-stream dense network. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR) **abs/1802.07412** (2018)
60. Zhang, H., Patel, V.M.: Image de-raining using a conditional generative adversarial network. arXiv preprint arXiv:1701.05957 (2017)
61. Zhang, H., Patel, V.M.: Densely connected pyramid dehazing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3194–3203 (2018)
62. Zhang, H., Sindagi, V., Patel, V.M.: Multi-scale single image dehazing using perceptual pyramid deep network. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 902–911 (2018)
63. Zhang, H., Sindagi, V., Patel, V.M.: Joint transmission map estimation and dehazing using deep networks. IEEE Transactions on Circuits and Systems for Video Technology (2019)
64. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2223–2232 (2017)
65. Zhu, X., Pang, J., Yang, C., Shi, J., Lin, D.: Adapting object detectors via selective cross-domain alignment. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 687–696 (2019)