Unified Density-Aware Image Dehazing and Object Detection in Real-World Hazy Scenes

Zhengxi Zhang, Liang Zhao, Yunan Liu, Shanshan Zhang(✉), and Jian Yang

PCA Lab, Key Lab of Intelligent Perception and Systems for High -Dimensional Information of Ministry of Education, Jiangsu Key Lab of Image and Video Understanding for Social Security, School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China
{zxzhang,liangzhao,liuyunan,shanshan.zhang,csjyang}@njust.edu.cn

Abstract. It is an important yet challenging task to detect objects on hazy images in real-world applications. The major challenge comes from low visual quality and large haze density variations. In this work, we aim to jointly solve the image dehazing and the object detection tasks in real hazy scenarios by using haze density as prior knowledge. Our proposed **U**nified **D**ehazing a**n**d **D**etection (UDnD) framework consists of three parts: a residual-aware haze density classifier, a density-aware dehazing network, and a density-aware object detector. First, the classifier exploits the residuals of hazy images to accurately predict density levels, which provide rich domain knowledge for the subsequent two tasks. Then, we design respectively a **H**igh-**R**esolution **D**ehazing **N**etwork (HRDN) and a Faster R-CNN-based multi-domain object detector to leverage the extracted density information and tackle hazy object detection. Experiments demonstrate that UDnD performs favorably against other methods for object detection in real-world hazy scenes. Also, HRDN achieves better results than state-of-the-art dehazing methods in terms of PSNR and SSIM. Hence, HRDN can conduct haze removal effectively, based on which UDnD is able to provide high-quality detection results.

1 Introduction

Object detection in hazy scenes is important for outdoor vision systems, *e.g.* video surveillance and autonomous driving; yet it is an extremely challenging task. The challenges mainly come from two aspects. On the one hand, hazy images are usually of poor visual quality that caused by low contrast, color distortion and blur *etc.* [1], making it more difficult to discriminate interesting objects from background clutters. On the other hand, haze density varies tremendously in real-world applications, leading to variations w.r.t. visual quality; these non-negligible intra-domain gaps make object detectors hard to converge.

Z. Zhang and L. Zhao—Equal contribution.

H. Ishikawa et al. (Eds.): ACCV 2020, LNCS 12625, pp. 119–135, 2021.
https://doi.org/10.1007/978-3-030-69538-5_8

A straightforward solution to hazy object detection is to first apply image dehazing and then perform object detection on dehazed images. Most previous work follow this strategy, isolating dehazing and detection [2,3]. Since dehazing methods are not able to fully recover latent clear images, it is not guaranteed that the dehazed images are optimal for object detection [2,4]. From this perspective, it is favorable to jointly solve the two tasks, so as to obtain detection-friendly dehazed images and more accurate detection results. In [5], a unified pipeline is first proposed for hazy object detection. However, in their method, each model is designed to process one fixed density level, without handling density variations.

Another line of work uses domain adaptation techniques to tackle the task. They take clear images as the source domain and hazy images as the target domain; and then they try to lift the target domain performance to the source domain level by closing the domain gap via feature alignment [6,7]. However, in practice the domain gap is too large to handle, and it becomes especially more complex when there even exist significant intra-domain gaps in the hazy domain.

In this paper, we deal with the above mentioned two challenges in one coherent framework by taking advantage of both lines of work. We perform image dehazing to reduce the clear-hazy domain gap and then use the simplest domain adaptation method of fine-tuning to adapt a detector based on the clear domain to the dehazed domain. In the whole procedure, we take into account the intra-domain differences of hazy images by separating feature extraction for different haze density levels. Specifically, we propose a **U**nified density-aware **D**ehazing a**n**d **D**etection (UDnD) framework for solving image dehazing and object detection in a joint way. First, a modified VGG-Net [8] is introduced to predict haze density using hazy residuals. Then, we design a density switch module to multiplex different haze levels. For dehazing, we make modifications to HRNetV2 [9] by up-sampling with transposed convolution [10] and summing up features from different scales. The object detector then takes the dehazed image as input and switches to the branch dictated by the density level.

The contributions of this work are as follows:

- We propose a UDnD framework to jointly solve dehazing and detection. It for the first time deals with the inter-domain and intra-domain gaps in both image dehazing and hazy object detection, making them mutually benefit.
- We build a residual-aware classifier that predicts haze density levels to assist image dehazing and object detection. To the best of our knowledge, we are the first to explicitly predict haze density as prior knowledge for Convolutional Neural Networks (CNNs).
- A novel dehazing method HRDN is introduced, which sums multi-resolution representations to recover finer details. Guided by haze levels, HRDN is able to integrate density-specific knowledge into the network so as to divide and conquer single image dehazing.
- Experiments are conducted on two real-world hazy datasets, where the proposed UDnD outperforms the vanilla detector and the density-unaware counterparts. We also evaluate our dehazing method on two synthetic datasets, showing better performance than previous state-of-the-art methods. These

results demonstrate that our unified framework can handle the two types of domain gaps and give more accurate detection results in real hazy conditions.

2 Related Work

Since we address the problem of hazy object detection by unifying single image dehazing and multi-domain learning, we will review related work in the above three aspects, respectively.

2.1 Hazy Object Detection

The performance of object detection has been boosted by deep learning. Many CNN-based detectors have been proposed during the past few years, including Faster R-CNN (FRCNN) [11], FPN [12], YOLO [13] and SSD [14]. Albeit obtaining satisfactory performance under clear-weather conditions, none of these models could work seamlessly in hazy scenes without some kind of adaptation.

An intuitive idea for solving hazy object detection is to adopt a two-stage approach, *i.e.* performing dehazing and detection separately. Following this, Li *et al.* [2] study the effect of dehazing on various detectors. They find that applying image dehazing as pre-processing is not very helpful and sometimes even harms the performance. In [3] and [4], similar conclusions are drawn for semantic segmentation and image classification. The main reason is that existing dehazing methods are not good enough to reconstruct high-quality clear images for subsequent high-level vision tasks [4]. To address this issue, Li *et al.* [5] jointly optimize dehazing and detection, achieving better results than traditional two-stage approaches on synthetic images. Though our method is also trained on synthetic images, we demonstrate end-to-end performance on real-world data, and our haze-density-specific gating function improves on their results.

On the other side, some methods adapt a detector from the clear domain to the hazy domain for hazy object detection. They typically find a way to measure the distance between feature distributions of both domains and then train a feature extractor to minimize that distance. Inspired by [15], recent work measure the distance by learning a domain classifier in an adversarial manner [6,7,16–19]. Chen *et al.* [6] present a Domain Adaptive Faster R-CNN (DA-FRCNN) to tackle image-level and instance-level domain shifts. [7] proposes to align the features from regions with objects. However, they do not consider intra-domain gaps in the target hazy domain, which are induced by density variations.

Our unified density-aware framework integrates ideas from both sides. The dehazing and the detection sub-networks are jointly optimized. Particularly, we alleviate the intra-domain gaps by utilizing density levels. In accordance with previous methods, we use FRCNN as the baseline detector for experiments. But in principle, our method can be applied to any arbitrary CNN-based detector.

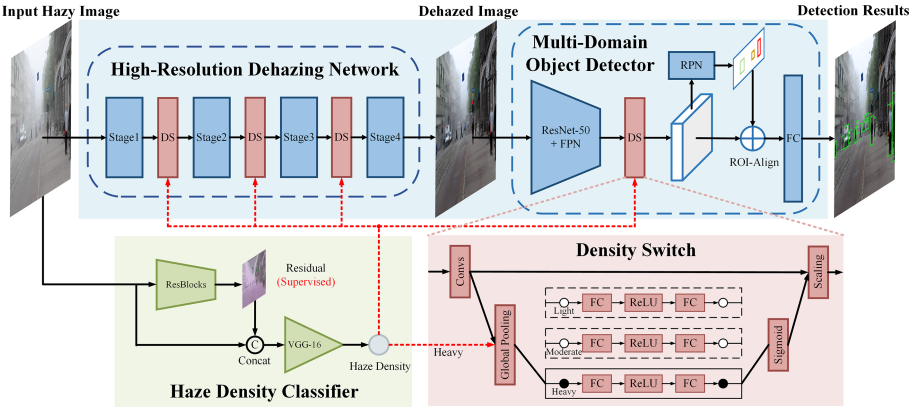


Fig. 1. An overview of the proposed unified density-aware image dehazing and object detection framework. It takes a real-world hazy image as input and first predicts its haze density level via a residual-aware classifier. The predicted density is then fed into a density switch module, which is used for multi-domain learning in the subsequent tasks. The whole network is optimized end-to-end.

2.2 Single Image Dehazing

Early dehazing methods stick to the standard optical model [20] and rely on hand-crafted priors [21–26]. Instead of manually designing features, CNN-based methods learn mappings directly from synthetic data. They usually estimate the transmission map and the atmospheric light, separately or jointly, as intermediate results, and then apply the reverse of the optical model [5,27–35]. However, estimating transmission in hazy scenes is an ill-posed problem, and it gets even worse when the colors of objects are similar to those of atmospheric lights [36]. Therefore, some methods [36–38] try to recover haze-free images directly via end-to-end frameworks, without reliance on the optical model.

The intra-domain gaps, *i.e.* haze density variations, cannot be ignored [3]. Some efforts have been made to incorporate haze density analysis into dark channel prior [39,40]. Dai *et al.* [41] train an AlexNet [42] to regress the attenuation coefficient. Recently, [30] uses multiple network stages to progressively estimate the transmission map and fuses the outputs from different stages, each of which is supervised by synthetic transmission of a fixed density level.

Instead, we handle the intra-domain gaps by explicitly predicting the haze density level and using it as prior knowledge for our end-to-end dehazing network.

2.3 Multi-Domain Learning

Multi-domain learning refers to learning effective representations for data from distinct domains [43]. It can be achieved by setting shared and domain-specific parameters, which resembles domain adaptation [44,45]. Previous work build domain-specific Batch Normalization (BN) layers [46] on otherwise shared networks [47–49]. Inspired by Squeeze-and-Excitation (SE) networks [50,51] introduces a data-driven SE adapter to adjust network activations.

In this work, we consider different haze levels as distinct domains and propose a density switch module to recalibrate features based on the haze density.

3 Proposed Method

In this section, we will first provide an overview of our proposed method and then explain each component in more detail.

3.1 Overview

We propose a coherent framework UDnD to jointly optimize image dehazing and object detection. Our method consists of three parts: a haze density classifier f, a dehazing module DH and a detection module DT. The classifier assigns each hazy image x^h a density level $\hat{d} = f(x^h)$. The dehazing module maps x^h to the latent clear image $\hat{x^c} = DH(x^h, \hat{d}; \theta_{DH})$, with $\hat{d}$ as domain knowledge. The detector takes $\hat{x^c}$ and $\hat{d}$ as input and outputs a structured prediction $\hat{y} = DT(\hat{x^c}, \hat{d}; \theta_{DT})$. Overall, our pipeline can be formulated as

$$\hat{y} = DT(DH(x^h, f(x^h); \theta_{DH}), f(x^h); \theta_{DT}), \tag{1}$$

where x^h is the hazy image and $\hat{y}$ is the detection result. We only presume DH and DT to be differentiable and assume nothing about f beyond providing discrete labels. The entire architecture is illustrated in Fig. 1.

Let x^h be a hazy image from the training set, with clear ground truth x^c and object detection annotations y. The overall loss function for our UDnD is

$$\begin{aligned} \mathcal{L}(x^h, x^c, y; \theta_{DH}, \theta_{DT}) = \lambda \mathcal{L}_{dehazing}(x^c, DH(x^h, f(x^h); \theta_{DH})) \\ + \mu \mathcal{L}_{detection}(y, DT(\hat{x^c}, f(x^h); \theta_{DT})), \end{aligned} \tag{2}$$

where $\hat{x^c} = DH(x^h, f(x^h); \theta_{DH})$ is the dehazed result of x^h. We use two weights λ and μ to balance the reconstruction term ($\mathcal{L}_{dehazing}$) and the task-driven term ($\mathcal{L}_{detection}$), which are described in Sect. 3.3 and Sect. 3.4 respectively. Note that the term $\hat{x^c}$ guarantees that the dehazing sub-network is supervised by the detection loss as long as μ is non-zero, while the dehazing loss does not directly affect the detection sub-network. The haze density classifier is used for extracting prior knowledge in our settings, thereby not updated in Eq. 2; which means our framework is compatible to prior-based density estimation methods [52] as well.

3.2 Residual-Aware Haze Density Classifier

The standard optical model [20] formulates the hazing process as

$$x^h(i) = x^c(i)t(i) + L(1 - t(i)), \tag{3}$$

where $x^h(i)$ is the observed hazy image at pixel location i, $x^c(i)$ is the clear scene radiance, and L is the atmospheric light. The transmission map $t(i)$ is obtained

using the distance $\ell(i)$ from the scene to the camera lens by $t(i) = \exp(-\beta\ell(i))$. Larger attenuation coefficient β indicates denser haze. For homogeneous haze, the Meteorological Optical Range (MOR) [53], *i.e.* visibility in meters, depends on β through $MOR = \frac{2.996}{\beta}$. It follows that $\beta \geq 2.996 \times 10^{-3}\,\mathrm{m}^{-1}$, where the equality holds for the lightest haze by definition.

We formulate haze density estimation as a classification problem. The predicted density should satisfy $\hat{d} \in \{1, \ldots, C\}$, where C is the total number of predefined density levels. Following [41], we set three levels: light, moderate and heavy; but our method can be extended to finer granularity given proper datasets. The haze density serves as domain label, guiding the update of domain-specific parameters in the subsequent dehazing and detection networks.

Inspired by [54], we observe that the residual of a hazy image, *i.e.* difference from its clear counterpart, is informative because the hazy image is a weighted sum of the clear image and the atmospheric light according to Eq. 3. Therefore, we propose a residual-aware haze density classifier that exploits the hazy residual. The details are depicted in Fig. 1. We stack 3 residual blocks [55] to estimate the residual, which is concatenated with the original hazy image to yield the 6-channel input for a modified VGG-16 [8].

Loss Function. The density classifier is optimized through a joint loss function:

$$\mathcal{L}_{classification} = \alpha\mathcal{L}_{res} + \mathcal{L}_{cls}, \tag{4}$$

where $\mathcal{L}_{res}$ is the L_1 loss for residual regression and $\mathcal{L}_{cls}$ is the cross-entropy loss for density classification. Additionally, α is used to balance the two tasks and is set to 0.2 in our experiments.

3.3 Density-Aware High-Resolution Dehazing Network

As is illustrated in Fig. 1, we propose a density-aware **H**igh-**R**esolution **D**ehazing **N**etwork (HRDN). The backbone is based on HRNetV2 [9], which maintains high-resolution representations and conducts repeated fusion to encourage interaction between multi-scale features. Different from HRNetV2, we only use 4 basic residual blocks in each network stage to prevent overfitting. The down-sampling operations in the stem are removed because dehazing is a pixel-level dense regression task. Moreover, we replace all the bilinear up-sampling units with transposed convolutions [10] to recover more details. To enforce a coarse-to-fine reconstruction process, we fuse the up-sampled features by summing them up instead of performing channel-wise concatenation like HRNetV2, which also reduces the computation cost as a side effect.

Density Switch. With the predicted haze density level from Sect. 3.2 as prior knowledge, we expect to handle density variations via multi-domain learning. Inspired by [51], we design a density switch module with multiple SE adapters, each corresponding to one type of density. From Fig. 1 we can see, the estimated haze level controls density switches by specifying which branch to take and what parameters to update, thus separating the feature extraction for different

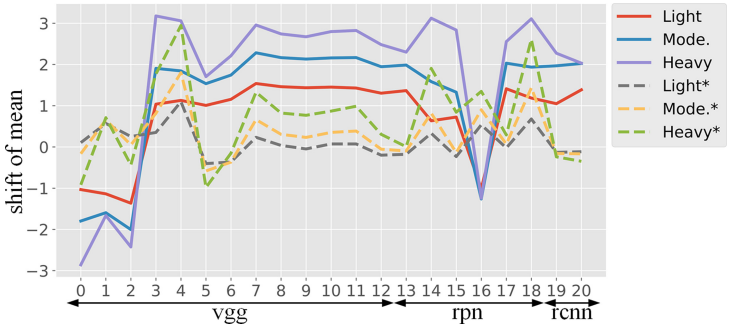


Fig. 2. Shift of mean values over convolutional activations of a vanilla FRCNN on various haze density levels. The vertical axis shows the difference between hazy (dehazed) and clear images. The horizontal axis gives the layer index. "*" indicates the dehazed images produced by our HRDN. The detector is trained on clear images. After dehazing, the activations become more similar to those of clear images, but certain intra-domain gaps remain.

densities. We add density switches before the 2^{nd}, 3^{rd} and 4^{th} stages of HRDN, enabling the network to divide and conquer the intra-domain gaps.

Loss Function. Existing dehazing methods utilize various loss functions, such as L_1 loss [37], MSE loss [5], smooth L_1 loss [31], perceptual loss [35], and adversarial loss [38]. Their weighted combinations are widely adopted. Despite improvements in performance, complicated loss functions increase the burden of hyper-parameter tuning and make the model hard to converge. Inspired by [56], we empirically find that a single SSIM loss works well:

$$\mathcal{L}_{dehazing}(x^c, \hat{x^c}) = 1 - SSIM(x^c, \hat{x^c}), \tag{5}$$

where x^c denotes the ground truth clear image and $\hat{x^c}$ denotes the dehazed image. The constant 1 here is added to ensure the loss value is non-negative.

3.4 Density-Aware Multi-Domain Object Detector

Although the hazy images have been processed by our dehazing network to reduce the inter-domain gaps, there still exist non-negligible intra-domain gaps among the dehazed images, which are caused by haze density variations. We provide some evidence via observing the convolutional activations of a vanilla FRCNN detector on the validation set of Foggy Cityscapes-DBF [41]. We collect the mean activations [51] for images of different densities, compute their differences from those of clear images before and after dehazing, and take these differences as domain gap measurements. A comparison is shown in Fig. 2. We have the following observations: (1) Prior to dehazing, the inter-domain gaps increase monotonically with haze density. (2) The inter-domain gaps are significantly reduced by dehazing, and thus dehazing serves as an effective pre-processing

step. (3) Even after dehazing, the intra-domain gaps remain for images of different density levels. These gaps need to be handled by the object detector. (4) The differences vary across layers. The first layers, which learn basic feature detection filters such as edges and corners, exhibit considerable amount of shifts. In other words, the domain gaps are not properly handled at the very beginning and they propagate forward, resulting in the final poor detection results.

In this work, we address the intra-domain gaps in the dehazed domain via multi-domain learning. Following [5,6], we make modifications on FRCNN. As is shown in Fig. 1, we introduce a density-aware multi-domain object detector by appending a density switch module to the ResNet-50 [55] and FPN [12]. By using the density switch, the detector will route images of different densities to desired branches, where different channel weights are computed to adjust the features. The weighted features are then fed into the Region Proposal Network (RPN) with density-specific information encoded.

Loss Function. We employ ROI-Alignment [57] to obtain the corresponding feature vector for each proposal from RPN. Finally, the category label is predicted via an ROI-wise classifier. The loss function for our multi-domain detector is inherited from the vanilla FRCNN [11] for simplicity:

$$\mathcal{L}_{detection} = \mathcal{L}_{rpn} + \mathcal{L}_{roi}. \tag{6}$$

Both the RPN loss ($\mathcal{L}_{rpn}$) and the ROI loss ($\mathcal{L}_{roi}$) consist of classification and localization terms, which are cross-entropy loss and smooth L_1 loss, respectively.

4 Experiments

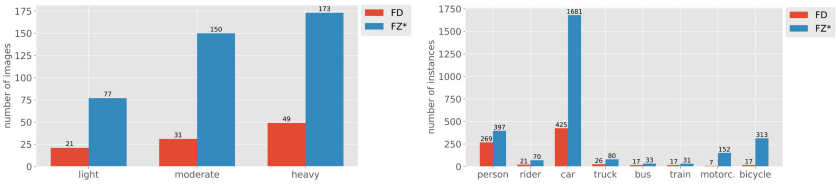
In this section, we first briefly introduce the datasets and the evaluation metrics used for our experiments, followed by the implementation details. After that, we will provide a comparison against other methods on hazy object detection to demonstrate the effectiveness of our UDnD framework. We will also show the performance of the proposed HRDN compared with state-of-the-art dehazing methods. Finally, we will do some ablation study in terms of domain adaptation techniques, loss functions, and unified training strategies.

4.1 Datasets

The object detectors are trained on synthetic hazy images generated using Eq. 3, but are evaluated on real hazy images. Whereas the dehazing methods are evaluated on synthetic data, only for which the ground truth are available.

Synthetic Datasets

OTS and SOTS-outdoor. RESIDE [2] contains both indoor and outdoor hazy scenes. We adopt the Outdoor Training Set (OTS) and the outdoor subset of Synthetic Objective Testing Set (SOTS-outdoor), and ensure the ground truth



Density distribution Class distribution

Fig. 3. Haze density distributions and class distributions of Foggy Driving (FD) and Foggy Zurich-test* (FZ*). Both datasets are composed of real-world hazy images. The density levels in (a) are predicted by our residual-aware haze density classifier. In (b), the relatively low number of instances in classes except car and person is not a surprise because hazy weather discourages road traffic.

clear images in OTS do not overlap with those in SOTS-outdoor through data cleaning [31]. The cleaned OTS has 296,695 hazy images with the atmospheric light $L \in [0.8, 1.0]$ and the attenuation coefficient $\beta \in [0.04, 0.2]$, generated out of 8,477 clear images. SOTS-outdoor has 500 hazy images.

Foggy Cityscapes-DBF. Foggy Cityscapes-DBF (FC-DBF) [41] derives from Cityscapes [58] and consists of a large and diverse set of urban street hazy scenes. There are a total of 8,925 images for training and 1,500 images for validation, both equally divided into three density levels ($\beta \in \{0.005, 0.01, 0.02\}$). We follow the screening criteria in [3] and use the selected 1,650 (550×3) high-quality synthetic hazy images to fine-tune the object detectors. This dataset is denoted as FC-DBF-refine. The bounding box annotations of these hazy images are automatically inherited from their clear-weather counterparts.

Real-World Datasets

Foggy Driving. Foggy Driving (FD) [3] is a collection of 101 hazy images of driving scenes, among which 51 images are captured at various areas of Zurich by a cell phone camera and others are selected from the Web.

Foggy Zurich. Foggy Zurich [41] is comprised of 3,808 images that are video frames depicting hazy road scenes in Zurich and its suburbs. Different from FD, these images are collected with a GoPro Hero 5 camera. We manually select 400 images of diverse scenes and haze densities, and annotate them carefully to create a new test set, namely Foggy Zurich-test* (FZ*). The statistics of FD and FZ* are shown in Fig. 3. We can see they both include various haze density levels. In particular, FZ* has significantly more annotated objects than FD, and thus can be served as a more convincing test set.

4.2 Evaluation Metrics

For hazy object detection, we adopt Average Precision (AP) and mean Average Precision (mAP) that is the average of APs over all classes. Additionally, the

Table 1. Comparison of different hazy object detection methods on Foggy Driving (FD) and Foggy Zurich-test* (FZ*) w.r.t. mAP (%). For training, "Clear" is clear-weather Cityscapes, "Syn. Hazy" is FC-DBF-refine , and "Real Hazy" is a subset of unlabelled Foggy Zurich. DA, UT and DL denote domain adaptation, unified training and density levels, respectively. Bold indicates the best results.

Methods	Components			Training sets			Test sets	
	DA	UT	DL	Clear	Syn. Hazy	Real Hazy	FD	FZ*
Vanilla FRCNN [11]				✓			18.26	17.31
DA-FRCNN [6]	✓			✓	✓		19.13	17.80
				✓		✓	19.58	18.52
				✓	✓	✓	22.68	19.27
JAOD-FRCNN [5]	✓	✓		✓	✓		22.17	19.14
Our UDnD	✓	✓	✓	✓	✓		**24.90**	**22.89**

mean of AP scores over the most frequent classes (car and person) is reported as mAP* for evaluating the detectors from a more practical perspective.

For dehazing, Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) [59] are used as standard quality measures. In particular, we investigate the effect of dehazing on detection with mAP.

4.3 Implementation Details

We take the FRCNN model pre-trained on Cityscapes by MMDetection [60,61] as baseline and initialization. The dehazing sub-network is pre-trained on OTS. First, we implement an improved version of two-stage approaches by freezing the dehazing part of our UDnD framework. We name this pipeline Dehazing and Detection (DnD) and fine-tune it on FC-DBF-refine for 9 epochs. SGD algorithm [62] is employed with a mini-batch size of 1 and an initial learning rate of 0.001, which decays polynomially. For UDnD, the input images are randomly cropped and resized to 512×512. The dehazing and the detection sub-networks are jointly optimized based on the DnD model with the same strategy. We set $\lambda = 1$ and $\mu = 1$ in Eq. 2 via cross-validation.

For dehazing, the models are trained from scratch with 256×256 image patches. Adam [63] is used for optimization with a mini-batch size of 4. The initial learning rate is 0.0001. We train the models up to 100 epochs on FC-DBF and adopt the cosine annealing schedule [64]. In consistent with [31], we train the networks on OTS with a patch size of 240×240 for 10 epochs and decay the learning rate by half after every 2 epochs. **Our code and trained models are available at** https://github.com/xiqi98/UDnD.

4.4 Comparison on Object Detection in Real-World Hazy Scenes

We choose three methods for comparison: vanilla FRCNN [11], a baseline trained on clear images only; DA-FRCNN [6], a state-of-the-art method using domain

Vanilla FRCNN [11] DA-FRCNN [6] JAOD-FRCNN [5] Our UDnD

Fig. 4. Examples of object detection results in real-world hazy scenes. Green boxes are true positives and red ones are missing recalls. In addition to detection, JAOD-FRCNN and the proposed UDnD provide dehazed images as well. JAOD-FRCNN tends to over-dehaze and produce artifacts that affect detection, while the vanilla FRCNN and DA-FRCNN miss some difficult objects. Our UDnD achieves a higher recall for hazy object detection.

adaptation; and JAOD-FRCNN [5], the only previous method jointly solving image dehazing and object detection.

We conduct evaluations on FD and FZ*, and report the results in Table 1, from which we have the following observations: (1) The vanilla FRCNN obtains the lowest mAP on both test sets, as it is trained on clear images only, without access to hazy images at all. It indicates that involving hazy images for training helps. (2) It boosts the performance by adapting the detector from clear to hazy images. For DA-FRCNN, the result on FD improves by more than 4 points over the baseline, when both synthetic and real-world data are used for training. Please note that for domain adaptation methods, real hazy images are needed to reach high performance. (3) JAOD-FRCNN achieves comparable performance to DA-FRCNN, showing that dehazing also serves as an effective way to reduce the inter-domain gaps. (4) Our method UDnD, which incorporates haze density information, obtains the best results. Specifically, it outperforms the vanilla FRCNN by $\sim$6 and $\sim$5 points on FD and FZ*, respectively; and it also improves on the results of DA-FRCNN, showing the benefits of a unified pipeline. Compared to JAOD-FRCNN, the obtained $\sim$3 points gain on FZ* demonstrates that the density switch module is helpful for dealing with the intra-domain differences, and thus leads to better detection performance.

Table 2. Comparison with state-of-the-art dehazing methods. To evaluate the effect of dehazing on object detection, we use a vanilla FRCNN based on Cityscapes to process the dehazed images produced by each model on OTS, and report mAP (%). Bold indicates the best results.

Methods	Dehazing				Detection	
	SOTS-outdoor		FC-DBF		FD	FZ*
	PSNR	SSIM	PSNR	SSIM	mAP	mAP
-/-	15.92	0.8029	16.07	0.8792	18.26	17.31
DCP [23]	16.32	0.8007	17.91	0.8749	17.07	17.50
NLD [21]	18.07	0.8016	16.53	0.8595	14.24	12.56
AOD-Net [5]	23.49	0.9063	20.79	0.9028	18.32	16.44
DCPDN [35]	26.74	0.9393	27.05	0.9630	19.79	17.71
EPDN [38]	29.61	0.9582	32.00	0.9812	19.25	18.62
GDN [31]	30.87	0.9832	33.07	0.9880	20.03	18.69
FFA-Net [37]	32.15	0.9806	30.99	0.9803	18.79	17.08
Our HRDN	**33.27**	**0.9877**	**35.08**	**0.9899**	**20.56**	**19.03**

Table 3. Effect of different domain adaptation techniques w.r.t. mAP and mAP* (%) that is the mean of APs over car and person. These experiments are conducted using our DnD pipeline. Bold indicates the best results.

Domain adaptation techniques			FD		FZ*	
Dehazing	Fine-tuning	Density switch	mAP	mAP*	mAP	mAP*
–	–	–	18.26	28.40	17.31	30.94
✓	–	–	19.84	29.29	18.26	32.45
✓	✓	–	22.70	29.68	19.57	34.85
✓	✓	✓	**23.15**	**30.42**	**20.49**	**35.33**

The qualitative comparison is illustrated in Fig. 4. We can see UDnD is better at handling small objects, occluded objects and dense haze, which demonstrates its effectiveness in real-world hazy object detection.

4.5 Comparison with State-of-the-Art Dehazing Methods

Our residual-aware haze density classifier achieves 97.60% accuracy on the validation set of FC-DBF, improving upon [41] by 3.33 points.

The proposed HRDN is evaluated against state-of-the-art dehazing methods [5,21,23,31,35,37,38] in terms of dehazing and detection performance. The results are shown in Table 2. We can observe that HRDN achieves the best performance on the two dehazing datasets, with margins of +1.12dB and +2.01dB in PSNR compared to the second best methods on SOTS-outdoor and the validation set of FC-DBF, respectively.

Table 4. Comparison of loss functions on the validation set of FC-DBF. The density switch modules in HRDN are disabled. Bold indicates the best results.

Loss functions	PSNR	SSIM
L_1	32.94	0.9831
Smooth L_1	31.58	0.9772
SSIM	31.11	0.9868
L_1 + *SSIM*	**33.11**	**0.9872**

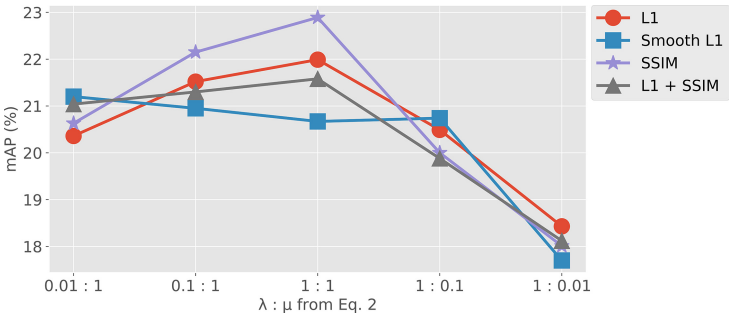


Fig. 5. Effect of different dehazing loss functions and $\lambda : \mu$ ratios on the detection performance of UDnD. We report mAP (%) on Foggy Zurich-test*.

Meanwhile, we use mAP as an additional task-driven metric. Hazy images from FD and FZ* are pre-processed by each dehazing model trained on OTS before fed into the vanilla FRCNN. From the last two columns of Table 2, we can see our HRDN obtains the best results w.r.t. mAP on both test sets. Some methods, *e.g.* FFA-Net, tend to overfit the synthetic training data, and thus obtain relatively low mAP on the two real-world datasets; other methods like DCPDN fall short on image dehazing.

To summarize, our dehazing method HRDN not only outperforms previous methods in terms of PSNR and SSIM, but is also more helpful for high-level vision tasks, such as object detection.

4.6 Ablation Study

Domain Adaptation. We study three techniques for tackling domain gaps, namely dehazing, fine-tuning and density switch, based on our DnD pipeline. Table 3 shows the effect of the three sequentially. Dehazing and fine-tuning are used to deal with the clear-hazy gaps, and the density switch module is used to handle haze density variations. They all bring performance gains of 1–2 points each, justifying that UDnD can handle both types of domain gaps.

Loss Function. In Fig. 5, we study the effect of different dehazing loss functions on the final detection performance, and find that a single SSIM loss works well.

Table 5. Effect of updating different sub-networks of UDnD. We report mAP and APs (%) over all classes of Foggy Zurich-test*. Bold indicates the best results.

Sub-Networks		pers.	rider	car	truck	bus	train	moto.	bicy.	mAP
Dehazing	Detection									
–	✓	25.34	21.63	45.32	3.26	14.46	30.07	13.37	10.48	20.49
✓	–	25.98	18.13	45.26	3.22	**16.93**	32.73	14.94	11.23	21.05
✓	✓	**26.03**	**24.41**	**48.66**	**5.74**	6.02	**37.00**	**21.46**	**13.78**	**22.89**

However, when we investigate on the dehazing task with the density switches in HRDN disabled, we observe that a combination of L_1 loss and SSIM loss achieves the best results, as is shown in Table 4. It indicates that there is a mismatch between image dehazing and object detection w.r.t. optimization goal, which explains why traditional two-stage methods fail in hazy object detection. We argue that haze mainly affects color, resulting in the wide application of L_1 loss and MSE loss in dehazing methods; but when it comes to detection, structure matters more than color. Hence, SSIM loss stands out as a good objective function for both tasks because of its emphasis on the structural information.

Unified Training. We evaluate our unified training strategy by freezing the weights of different pre-trained sub-networks while keeping the same loss. The results in Table 5 show that disabling joint optimization of the dehazing and the detection sub-networks leads to a performance drop of ∼2 points w.r.t. mAP, manifesting the importance of a unified framework.

5 Conclusion

We have presented a **U**nified density-aware **D**ehazing a**n**d **D**etection (UDnD) framework for image reconstruction and object detection in hazy conditions, motivated by the ideas to jointly optimize the two tasks and to exploit haze density as prior knowledge. We propose a residual-aware classifier to estimate haze density, a density-aware **H**igh-**R**esolution **D**ehazing **N**etwork (HRDN) to divide and conquer various hazy scenarios, and a density-aware multi-domain object detector to tackle the final detection task. These collectively constitute a unified pipeline for hazy object detection. Experiments demonstrate the effectiveness of each module and the entire framework in real-world hazy scenes.

Acknowledgement. This work was supported by the National Science Fund of China (Grant Nos. 61702262, U1713208), Funds for International Cooperation and Exchange of the National Natural Science Foundation of China (Grant No. 61861136011), Natural Science Foundation of Jiangsu Province, China (Grant No. BK20181299), Young Elite Scientists Sponsorship Program by CAST (2018QNRC001), the Fundamental Research Funds for the Central Universities" (Grant No.30920032201), and Science and Technology on Parallel and Distributed Processing Laboratory (PDL) Open Fund (WDZC20195500106).

References

1. Li, Y., You, S., Brown, M.S., Tan, R.T.: Haze visibility enhancement: a survey and quantitative benchmarking. CVIU **165**, 1–16 (2017)
2. Li, B., et al.: Benchmarking single-image dehazing and beyond. IEEE Trans. Image Process. **28**, 492–505 (2019)
3. Sakaridis, C., Dai, D., Van Gool, L.: Semantic foggy scene understanding with synthetic data. IJCV **126**, 973–992 (2018)
4. Pei, Y., Huang, Y., Zou, Q., Lu, Y., Wang, S.: Does haze removal help CNN-based image classification? In: ECCV, pp. 682–697 (2018)
5. Li, B., Peng, X., Wang, Z., Xu, J., Feng, D.: Aod-net: all-in-one dehazing network. In: ICCV, pp. 4770–4778 (2017)
6. Chen, Y., Li, W., Sakaridis, C., Dai, D., Van Gool, L.: Domain adaptive faster R-CNN for object detection in the wild. In: CVPR, pp. 3339–3348 (2018)
7. Zhu, X., Pang, J., Yang, C., Shi, J., Lin, D.: Adapting object detectors via selective cross-domain alignment. In: CVPR, pp. 687–696 (2019)
8. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015)
9. Wang, J., et al.: Deep high-resolution representation learning for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. (2019)
10. Dumoulin, V., Visin, F.: A guide to convolution arithmetic for deep learning. arXiv preprint arXiv:1603.07285 (2016)
11. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: towards real-time object detection with region proposal networks. TPAMI **39**, 1137–1149 (2017)
12. Lin, T.Y., Dollar, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: CVPR, pp. 2117–2125 (2017)
13. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: CVPR, pp. 779–788 (2016)
14. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: SSD: single shot multibox detector. In: ECCV, pp. 21–37 (2016)
15. Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Vaughan, J.W.: A theory of learning from different domains. Mach. Learn. **79**, 151–175 (2010)
16. He, Z., Zhang, L.: Multi-adversarial faster-RCNN for unrestricted object detection. In: ICCV, pp. 6668–6677 (2019)
17. Saito, K., Ushiku, Y., Harada, T., Saenko, K.: Strong-weak distribution alignment for adaptive object detection. In: CVPR, pp. 6956–6965 (2019)
18. Saito, K., Watanabe, K., Ushiku, Y., Harada, T.: Maximum classifier discrepancy for unsupervised domain adaptation. In: CVPR, pp. 3723–3732 (2018)
19. Zheng, Y., Huang, D., Liu, S., Wang, Y.: Cross-domain object detection through coarse-to-fine feature adaptation. arXiv preprint arXiv:2003.10275 (2020)
20. Koschmieder, H.: Theorie der horizontalen sichtweite. Beitrage zur Physik der freien Atmosphare, pp. 33–53 (1924)
21. Berman, D., treibitz, T., Avidan, S.: Non-local image dehazing. In: CVPR, pp. 1674–1682 (2016)
22. Fattal, R.: Dehazing using color-lines. TOG **34**, 1–14 (2014)
23. He, K., Sun, J., Tang, X.: Single image haze removal using dark channel prior. TPAMI **33**, 2341–2353 (2011)
24. Meng, G., Wang, Y., Duan, J., Xiang, S., Pan, C.: Efficient image dehazing with boundary constraint and contextual regularization. In: ICCV, pp. 617–624 (2013)
25. Tan, R.T.: Visibility in bad weather from a single image. In: CVPR, pp. 1–8 (2008)

26. Zhu, Q., Mai, J., Shao, L.: A fast single image haze removal algorithm using color attenuation prior. TIP **24**, 3522–3533 (2015)
27. Cai, B., Xu, X., Jia, K., Qing, C., Tao, D.: Dehazenet: an end-to-end system for single image haze removal. TIP **25**, 5187–5198 (2016)
28. Deng, Z., et al.: Deep multi-model fusion for single-image dehazing. In: ICCV, pp. 2453–2462 (2019)
29. Li, R., Pan, J., Li, Z., Tang, J.: Single image dehazing via conditional generative adversarial network. In: CVPR, pp. 8202–8211 (2018)
30. Li, Y., et al.: Lap-net: level-aware progressive network for image dehazing. In: ICCV, pp. 3276–3285 (2019)
31. Liu, X., Ma, Y., Shi, Z., Chen, J.: Griddehazenet: attention-based multi-scale network for image dehazing. In: ICCV, pp. 7314–7323 (2019)
32. Liu, Y., Pan, J., Ren, J., Su, Z.: Learning deep priors for image dehazing. In: ICCV, pp. 2492–2500 (2019)
33. Ren, W., Liu, S., Zhang, H., Pan, J., Cao, X., Yang, M.H.: Single image dehazing via multi-scale convolutional neural networks. In: ECCV, pp. 154–169 (2016)
34. Yang, D., Sun, J.: Proximal dehaze-net: a prior learning-based deep network for single image dehazing. In: ECCV, pp. 702–717 (2018)
35. Zhang, H., Patel, V.M.: Densely connected pyramid dehazing network. In: CVPR, pp. 3194–3203 (2018)
36. Ren, W., et al.: Gated fusion network for single image dehazing. In: CVPR, pp. 3253–3261 (2018)
37. Qin, X., Wang, Z., Bai, Y., Xie, X., Jia, H.: Ffa-net: feature fusion attention network for single image dehazing. In: AAAI, pp. 11908–11915 (2020)
38. Qu, Y., Chen, Y., Huang, J., Xie, Y.: Enhanced pix2pix dehazing network. In: CVPR, pp. 8160–8168 (2019)
39. Li, R., Kintak, U.: Haze density estimation and dark channel prior based image defogging. In: ICWAPR, pp. 29–35 (2018)
40. Yeh, C.H., Kang, L.W., Lin, C.Y., Lin, C.Y.: Efficient image/video dehazing through haze density analysis based on pixel-based dark channel prior. In: ISIC, pp. 238–241 (2012)
41. Dai, D., Sakaridis, C., Hecker, S., Van Gool, L.: Curriculum model adaptation with synthetic and real data for semantic foggy scene understanding. IJCV **128**, 1182–1204 (2019)
42. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NeurIPS, pp. 1097–1105 (2012)
43. Nam, H., Han, B.: Learning multi-domain convolutional neural networks for visual tracking. In: CVPR, pp. 4293–4302 (2016)
44. Long, M., Cao, Y., Wang, J., Jordan, M.I.: Learning transferable features with deep adaptation networks. arXiv preprint arXiv:1502.02791 (2015)
45. Mallya, A., Davis, D., Lazebnik, S.: Piggyback: adapting a single network to multiple tasks by learning to mask weights. In: ECCV, pp. 67–82 (2018)
46. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: ICML, pp. 448–456 (2015)
47. Bilen, H., Vedaldi, A.: Universal representations: The missing link between faces, text, planktons, and cat breeds. arXiv preprint arXiv:1701.07275 (2017)
48. Rebuffi, S.A., Bilen, H., Vedaldi, A.: Learning multiple visual domains with residual adapters. In: NeurIPS, pp. 506–516 (2017)
49. Rebuffi, S.A., Bilen, H., Vedaldi, A.: Efficient parametrization of multi-domain deep neural networks. In: CVPR, pp. 8119–8127 (2018)

50. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: CVPR, pp. 7132–7141 (2018)
51. Wang, X., Cai, Z., Gao, D., Vasconcelos, N.: Towards universal object detection by domain attention. In: CVPR, pp. 7289–7298 (2019)
52. Choi, L.K., You, J., Bovik, A.C.: Referenceless prediction of perceptual fog density and perceptual image defogging. TIP **24**, 3888–3901 (2015)
53. NOAA: Federal meteorological handbook no. 1: Surface weather observations and reports (2005)
54. Zhang, H., Patel, V.M.: Density-aware single image de-raining using a multi-stream dense network. In: CVPR, pp. 695–704 (2018)
55. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2016)
56. Ren, D., Zuo, W., Hu, Q., Zhu, P., Meng, D.: Progressive image deraining networks: a better and simpler baseline. In: CVPR, pp. 3937–3946 (2019)
57. He, K., Gkioxari, G., Dollar, P., Girshick, R.: Mask R-CNN. In: ICCV, pp. 2961–2969 (2017)
58. Cordts, M., et al.: The cityscapes dataset for semantic urban scene understanding. In: CVPR, pp. 3213–3223 (2016)
59. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. TIP **13**, 600–612 (2004)
60. Chen, K., et al.: Mmdetection: open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019)
61. Paszke, A., et al.: Pytorch: an imperative style, high-performance deep learning library. In: NeurIPS, pp. 8024–8035 (2019)
62. Ruder, S.: An overview of gradient descent optimization algorithms. arXiv preprint arXiv:1609.04747 (2016)
63. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
64. Loshchilov, I., Hutter, F.: SGDR: stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)