Hindawi
International Journal of Intelligent Systems
Volume 2023, Article ID 3691730, 15 pages
https://doi.org/10.1155/2023/3691730

Research Article

HLA-HOD: Joint High-Low Adaptation for Object Detection in Hazy Weather Conditions

Yiyang Shen,[1] **Rongwei Yu**,[1] **Ni Shu**,[1] **Jing Qin**,[2] **and Mingqiang Wei**[3]

[1]*Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University, Wuhan, China*
[2]*The Hong Kong Polytechnic University, Hong Kong SAR, China*
[3]*School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China*

Correspondence should be addressed to Rongwei Yu; roewe.yu@whu.edu.cn

Received 10 December 2022; Revised 13 March 2023; Accepted 15 March 2023; Published 14 April 2023

Academic Editor: Vasudevan Rajamohan

Object detection remains challenging in hazy weather conditions due to the poor visibility of captured images. There are currently two types of detectors capable of adapting to varying weather conditions: (i) low-level adaptation methods that combine one detector with an additional dehazing network and (ii) high-level adaptation methods that explore various kinds of domain adaptation knowledge. However, neither of these approaches can achieve desirable performance due to their inherent limitations. We raise an intriguing question—if combining both low-level adaptation and high-level adaptation, can improve the generalization ability of a detector in hazy weather conditions? To answer it, we propose a Joint High-Low Adaptation Object Detection paradigm (HLA-HOD) in hazy weather conditions. By combining both low-level adaptation and high-level adaptation, HLA-HOD achieves superior performance on hazy images without requiring ground-truth bounding boxes or clean images. Extensive experiments demonstrate that our method outperforms state-of-the-art low-level and high-level adaptation methods by a large margin both quantitatively and qualitatively.

1. Introduction

Object detection is a fundamental computer vision task and has been deployed in many real-life applications [1, 2]. In recent years, a variety of object detection approaches have emerged and achieved promising results on normal images [3–7]. Object detection approaches can be broadly categorized into two main types: region proposal-based methods and regression-based methods. The region proposals with CNNs (R-CNN family) [8–10] belong to the first category. These two-stage approaches first generate regions of interest (RoIs) from the input image and subsequently classify them by training neural networks. The YOLO family [3–5], SSD family [11–14], and RetinaNet [15] represent one-stageregression-based approaches, where bounding box coordinates and class labels are directly predicted from images using a single CNN. However, these methods fail to achieve desirable results under adverse weather conditions, especially in the presence of haze, which is one of the most frequently occurring weather phenomena in driving scenarios. Images captured on such hazy days inevitably suffer from the noticeable degradation of visual quality, severely worsening the performance of object detection.

Recently, numerous approaches have been proposed to restore the clean images from the hazy images. Early approaches heavily rely on hand-crafted priors, such as dark channel prior [16], color attenuation prior [17], and contrast maximization [18]. However, these hand-crafted priors exhibit a limited representation capacity, resulting in poor dehazed results under complex and diverse haze scenarios. To rectify this weakness, deep learning-based methods have made great progress, utilizing structures such as CNN [19–22], GAN [23–25], and Transformer [26, 27] for image dehazing. However, these approaches, designed primarily for human visual perception, may not always be beneficial or applicable to complex computer vision tasks, such as object

detection. As a result, the issue of improving object detection accuracy in hazy weather conditions has garnered increasing attention in the computer vision community.

To tackle this challenging problem, it is reasonable to consider both the image dehazing task (low-level) and the object detection task (high-level). We observe that there are two types of prevailing learning-based paradigms: (i) low-level adaptation methods introduce existing image dehazing algorithms to reduce the low-level gap in pixel-level appearances, such as the haze degradation, camera noise, and color bias. They employ two subnetworks to jointly perform image dehazing and object detection in a cascaded manner [28–30] or a parallel manner [31–33], as shown in Figures 1(a) and 1(b), respectively; (ii) high-level adaptation methods aim to adapt the object detection model from normal images to hazy images by exploring special domain generalization priors such as alignment [34–36], adversarial learning [37–39], and pseudolabeling [40, 41], as shown in Figure 1(c).

However, both paradigms exhibit certain limitations: (i) low-level adaptation methods heavily rely on clean images and ground-truth bounding box annotations during training, which limits their robustness and flexibility in real-time or resource-constrained applications. Additionally, it is challenging to balance the weights between image dehazing and object detection; (ii) high-level adaptation methods do not require dehazing modules or labels of hazy images, but they face difficulties in adapting from normal to hazy images due to the huge gap between hazy and normal images. Additionally, they only utilize domain adaptation knowledge, ignoring weather-specific information, which makes them vulnerable to the effects of haze degradation on detection performance. To this end, we are motivated to raise an intriguing question—can we combine both low-level adaptation and high-level adaptation to enhance the object detection performance in hazy weather conditions?

In this paper, we consider combining low-level and high-level adaptation to propose a joint high-low adaptation object detection paradigm (HLA-HOD) in hazy weather conditions (see Figure 1(d)). HLA-HOD is a two-stage network: in the low-level adaptation stage, we leverage the physical properties of the real-world hazy environment [42, 43] to obtain the dehazed and rehazed results. These results serve as intermediate states for the high-level adaptation stage. In the high-level adaptation stage, multitask learning and contrastive learning techniques are utilized to push the feature spaces of dehazed, rehazed, and normal states towards each other, instead of directly closing the gap between the feature spaces of haze and normal states. By combining low-level and high-level adaptation, HLA-HOD outperforms the state-of-the-art methods even without ground-truth data. The main contributions are as follows:

(i) By combining low-level and high-level adaptation, we propose a novel joint high-low adaptation object detection paradigm (HLA-HOD) in hazy weather conditions. HLA-HOD outperforms both low-level and high-level adaptation methods without requiring ground-truth bounding boxes or clean images.

(ii) For the low-level adaptation stage, we employ weather-specific priors to construct a haze-clean--haze and a clean-haze-clean translation process, enabling the acquisition of dehazed and rehazed images without ground-truth images. These dehazed and rehazed images are subsequently collected and used to establish two intermediate states for high-level adaptation.

(iii) For the high-level adaptation stage, we leverage contrastive learning and multitask learning to effectively close the feature spaces of dehazed, rehazed, and normal domains, thus avoiding the failure of directly closing the feature spaces of normal and hazy states.

2. Methodology

The presence of haze often leads to significant visibility degradation, making it challenging for general object detection models to obtain satisfactory performance. To tackle this issue, it is reasonable to recall two common tasks in the field of computer vision, i.e., image dehazing (a low-level vision task) and object detection (a high-level vision task). Existing solutions can be roughly divided into two categories: low-level adaptation and high-level adaptation. Low-level adaptation methods attempt to combine an object detection network with an additional dehazing network in a cascaded or parallel manner. High-level adaptation methods utilize existing domain adaptation priors to adapt the detection model from the normal images to the hazy images. However, these methods still achieve suboptimal performance on hazy images. Beyond existing wisdom, we propose a novel joint high-low adaptation object detection paradigm in hazy weather conditions, referred to as HLA-HOD. As shown in Figure 2, HLA-HOD is a two-stage network: in the low-level adaptation stage, we leverage unpaired hazy and clean images to generate two intermediate states. These states are then utilized for high-level adaptation in the subsequent stage. In the high-level adaptation stage, we avoid directly closing the feature spaces of normal and hazy states. Instead, we explore multitask learning and contrast learning to push the feature spaces of multiple states towards each other and obtain the final detection results.

2.1. Low-Level Adaptation Stage. Prior low-level adaptation methods heavily rely on image dehazing techniques to enhance object detection performance. However, a significant limitation of these methods is their dependence on synthetic paired hazy images for training, due to the difficulty in collecting real-world hazy/clean image pairs. As a result, their generalization capability is limited, especially in real-time and resource-constrained applications. To tackle this issue, CycleGAN [44] designs a haze-clean-haze cycle translation to obtain dehazed results without requiring clean images. However, due to the complexity of the haze

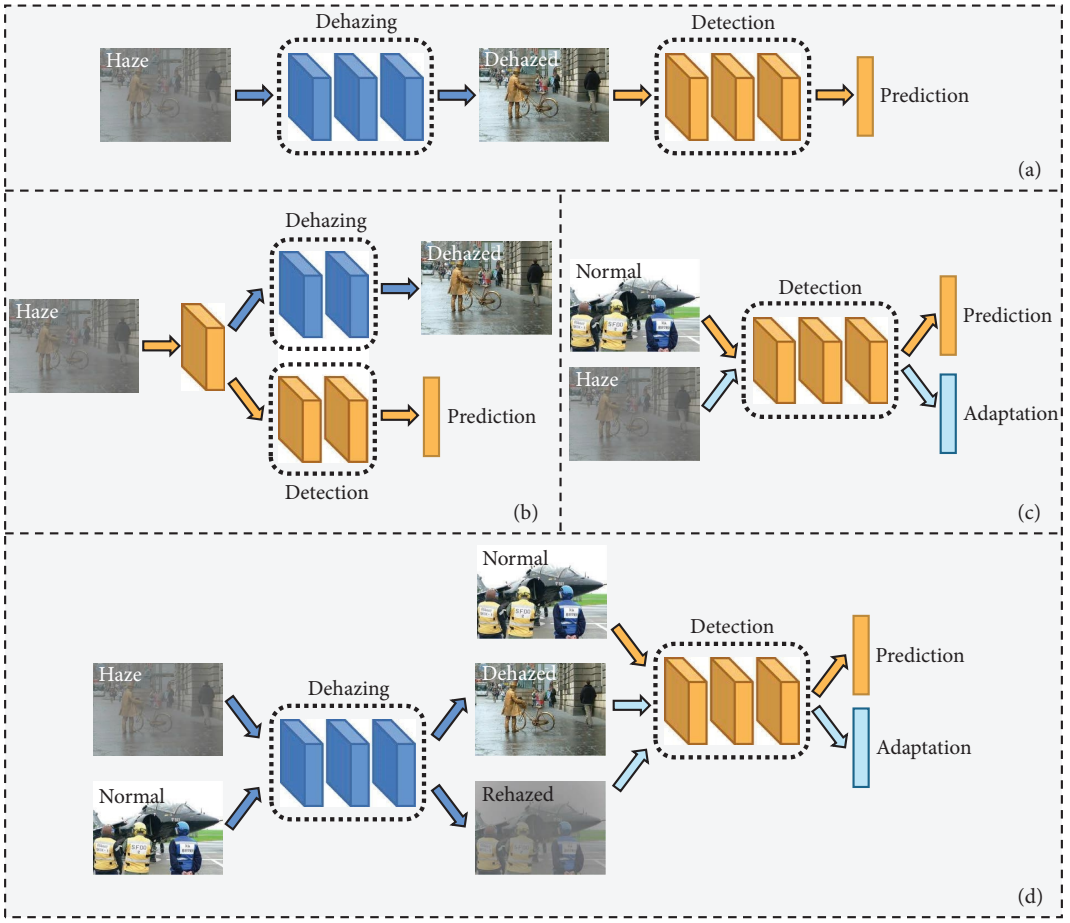


FIGURE 1: Illustration of four types of approaches for tackling the problem of object detection in the presence of haze. Low-level adaptation methods employ two subnetworks to jointly perform image dehazing and object detection in a cascaded manner (a) or a parallel manner (b) High-level adaptation methods employ domain adaptation priors to adapt the object detection model from normal images to hazy images (c) Beyond existing wisdom, we combine low-level and high-level adaptation techniques to propose a joint high-low adaptation object detection paradigm (HLA-HOD) in hazy weather conditions (d).

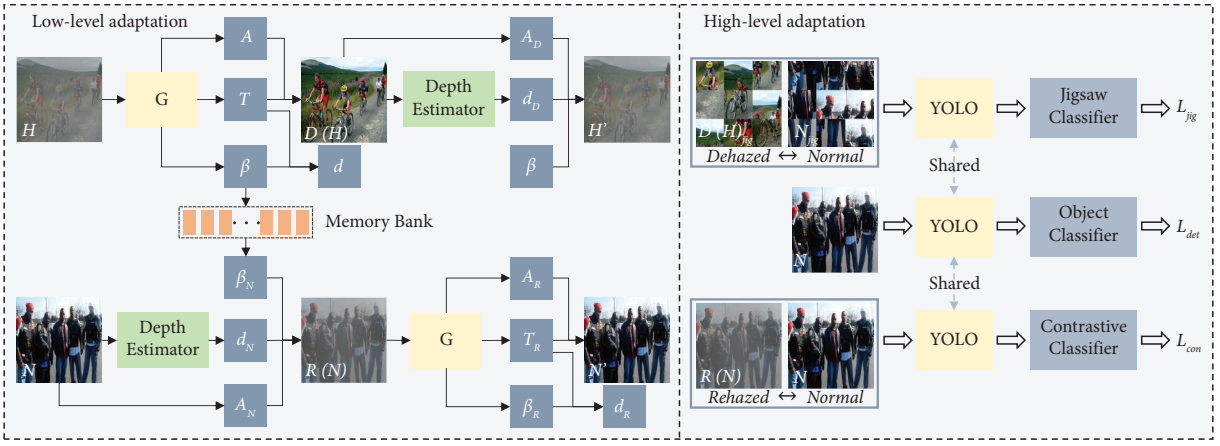


FIGURE 2: The pipeline of HLA-HOD. HLA-HOD is a two-stage network consisting of the low-level adaptation stage and the high-level adaptation stage. The low-level adaptation stage obtains two intermediate states for high-level adaptation. The high-level adaptation stage leverages multitask learning and contrastive learning to push the feature spaces of multiple states towards each other and obtain the final detection results. Remarkably, HLA-HOD does not require ground-truth bounding boxes or clean images.

degradation process, simple image-to-image translation methods such as CycleGAN and its variants often cause color and/or structure distortions, severely hindering the object detection performance.

To tackle these issues, we initially formulate a haze imaging model based on the physical properties of the real-world haze. In the low-level adaptation stage (LAS), we follow the established haze imaging model to decompose physical factors into three components: atmospheric light, transmission map, and scattering coefficient. It allows us to generate more satisfactory dehazed and rehazed results compared to CycleGAN. Furthermore, these results serve as two intermediate states for high-level adaptation, which facilitates the process of adapting from clean images to hazy images. LAS consists of two branches: (i) the haze-clean--haze branch $H \longrightarrow D(H) \longrightarrow H^{'}$ and (ii) the clean-haze-clean branch $N \longrightarrow R(N) \longrightarrow N^{'}$.

2.1.1. Hazy Image Formulation. According to [42, 43], the captured hazy image $I(x)$ at a pixel x can be formulated as

$$I(x) = B(x)t(x) + A(1 - t(x)), \tag{1}$$

where B and A represent the background and global atmosphere light, respectively. The transmission map $t(x)$ is defined as

$$t(x) = e^{-\beta d(x)}, \tag{2}$$

where $d(x)$ and β represent the scene depth and scattering coefficient, respectively.

2.1.2. Haze-Clean-Haze Branch. The hazy image H is fed into the generator to obtain the atmospheric light A, the transmission map T, and the scattering coefficient β, as shown in Figure 2. Notably, we utilize RefineNet [45] as the baseline to construct our generator, as depicted in Figure 3. We first divide the pretrained ResNet [46] into four blocks, which enable us to extract multiscale features with sizes of 1/4, 1/8, 1/16, and 1/32 of the original image. These multiscale features are then fed into four pool layers to obtain four pooling results $\beta_{i\in\{1,2,3,4\}}$, which are concatenated into a 3× 3 convolution layer followed by a normal convolution layer to obtain the final scattering coefficient β. We further incorporate a unique memory bank to record various scattering coefficient parameters from the generator.

To obtain transmission map T, we start from the last block of the ResNet and connect the output of the ResNet-4 block to RefineNet-4. Notably, RefineNet-4 only has a single input and serves as an additional set of convolutions. In the next stage, we feed the output of the ResNet-3 block into a self-attention layer [47] instead of standard convolution layers [45] since the self-attention mechanism can enhance CNNs with global interactions and address the limitations of repeating local operations. The outputs of Att-3 and RefineNet-4 are fed into RefineNet-3 as 2-path inputs. RefineNet-3 can utilize the high-resolution features from Att-3 to refine the low-resolution feature map output by RefineNet-4 in the previous stage. RefineNet-2 and RefineNet-1 are the same as RefineNet-3, which both fuse high-resolution features from the later layers and low-resolution features from the earlier ones. Finally, the high-resolution feature maps from RefineNet-1 are fed into a dense soft-max layer to obtain the prediction of the transmission map. Furthermore, we adopt the dark channel prior to obtaining atmospheric light A. Based on equations (1) and (2), we can compute both the dehazed result $D(H)$ and the depth map d.

To obtain the rehazed result $H^{'}$, we initially employ a depth estimator, with the same structure as the generator illustrated in Figure 3, to obtain the corresponding depth map d_D. Notably, we only retain the process of transmission map estimation to obtain the estimated depth map. Furthermore, we select the brightest pixel as the atmospheric light A_D to generate the haze. Given the depth map d_D, atmospheric light A_D, and the previously estimated scattering coefficient β, we can employ equations (1) and (2) to rehaze the dehazed image $D(H)$.

2.1.3. Clean-Haze-Clean Branch. The clean-haze-clean branch is similar to the haze-clean-haze branch, utilizing the same generator and depth estimator. Initially, we feed the unpaired clean image N into the depth estimator to obtain the corresponding depth map d_N. Additionally, A_N is also extracted from the brightest pixel. To improve the generalization ability in different haze densities, we randomly read the scattering coefficient β_N from the memory bank to generate hazy images with different haze densities. The synthetic hazy image $R(N)$ can be generated using equations (1) and (2). Following the same physical process, we then input $R(N)$ into the generator to obtain the depth map d_R and dehazed result $N^{'}$.

2.1.4. Loss Function. Different from most of the existing low-level adaptation methods, we comprehensively consider multiloss function to constrain the unsupervised training process and preserve the color and structure of images. The multiloss function includes adversarial loss, cycle-consistency loss, depth loss, and scattering coefficient loss. The formulation is as follows:

$$L_{\text{low-level}} = \lambda_{\text{adv}} L_{\text{adv}} + \lambda_{\text{cyc}} L_{\text{cyc}} + \lambda_{\text{dep}} L_{\text{dep}} + \lambda_{\text{sca}} L_{\text{sca}}, \tag{3}$$

where λ_{adv}, λ_{cyc}, λ_{dep}, and λ_{sca} are balanced weights that we experimentally set to 0.5, 1, 1, and 1, respectively.

We employ the adversarial loss to minimize the distance between the generated images and the target images. Regular GANs loss may lead to the vanishing gradient problem during the training process. Thus, we resort to the original Least-Squares GAN (LSGAN) [48] to serve as the adversarial loss:

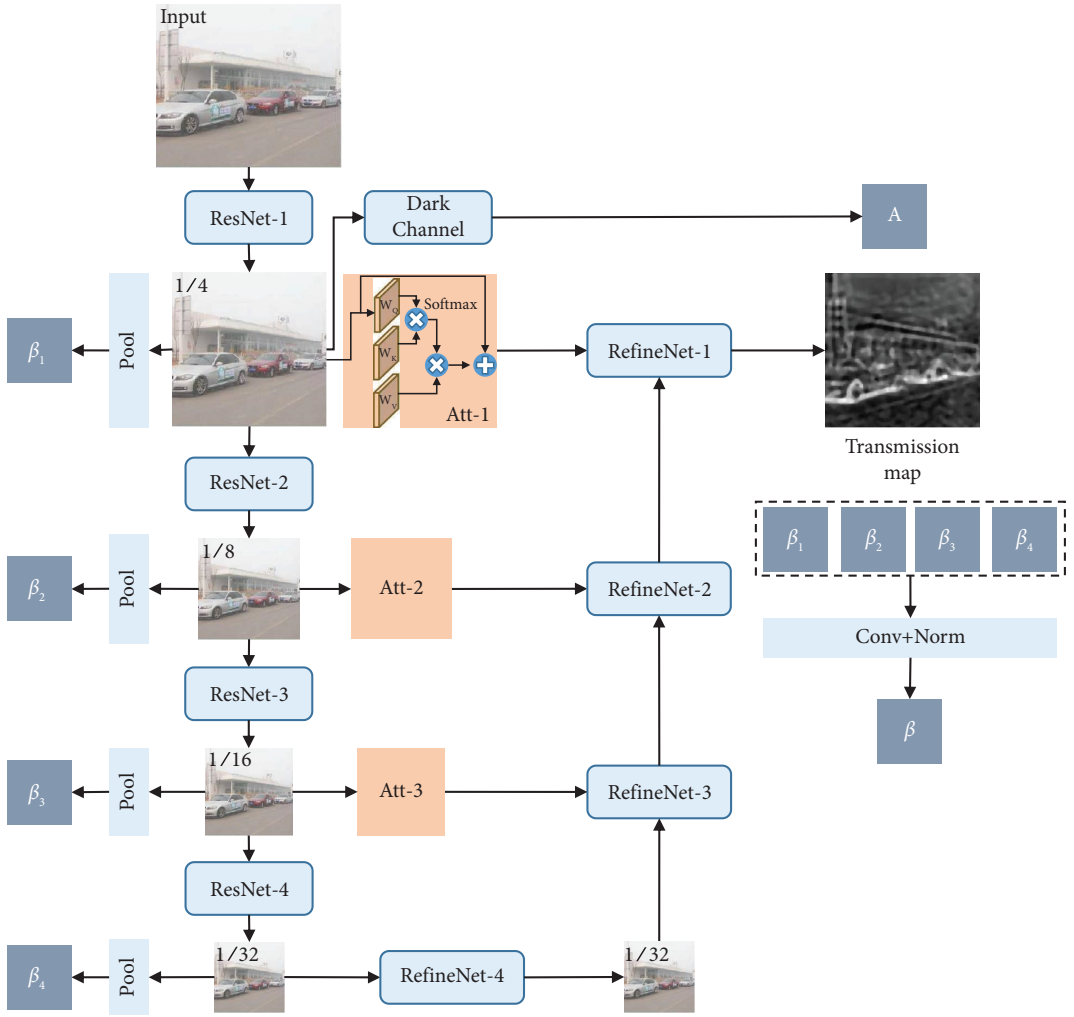


Figure 3: The architecture of the generator.

$$
\begin{aligned}
L_{\text{adv}}(G) &= E_{D(H)\sim P_{\text{fake}}}\left[(D_A(D(H))-1)^2\right] \\
&\quad + E_{R(N)\sim P_{\text{fake}}}\left[(D_B(R(N)-1)^2\right], \\
L_{\text{adv}}(D_A) &= E_{N\sim P_{\text{real}}}\left[(D_A(N)-1)^2\right] \\
&\quad + E_{D(H)\sim P_{\text{fake}}}\left[(D_A(D(H)))^2\right], \\
L_{\text{adv}}(D_B) &= E_{H\sim P_{\text{real}}}\left[(D_B(H)-1)^2\right] \\
&\quad + E_{R(N)\sim P_{\text{fake}}}\left[(D_B(R(N)))^2\right],
\end{aligned} \tag{4}
$$

where D_A and D_B represent discriminators [49].

We utilize the cycle-consistency loss to constrain the space of generated samples and retain the contents of generated images, defined as follows:

$$
L_{\text{cyc}} = E_{H\sim P_{\text{data}(H)}}\left[\left\|H - H^{'}\right\|_1\right] + E_{N\sim P_{\text{data}(N)}}\left[\left\|N - N^{'}\right\|_1\right], \tag{5}
$$

where $\|.\|_1$ is the $L1$ norm.

We adopt the depth loss to limit the consistency between depth maps generated by the generator and the depth estimator, which is formulated as

$$
L_{\text{dep}} = \left\|d - d_D\right\|_1 + \left\|d_N - d_R\right\|_1, \tag{6}
$$

where $\|.\|_1$ is the $L1$ norm.

We adopt the scattering coefficient loss to penalize the difference between β_N and β_R as

$$
L_{\text{sca}} = \left\|\beta_N - \beta_R\right\|_1, \tag{7}
$$

where β_N is randomly read from the memory bank to enhance the model's ability to generalize across diverse haze densities.

2.2. High-Level Adaptation Stage. Most of the high-level adaptation methods attempt to use special domain generalization priors, e.g., alignment, adversarial learning, and pseudolabeling, to adapt the detectors from normal to hazy images. However, these methods are insufficient in bridging the huge gap between normal and hazy images. To tackle this issue, we initially collect the dehazed images $D(H)$ and rehazed images $R(N)$ from the low-level adaptation stage to set up two intermediate states, i.e., dehazed and rehazed. Instead of directly closing $N \longleftrightarrow H$, we push the feature spaces of N, $D(H)$, and $R(N)$ towards each other.

Specifically, we employ cross-domain multitask learning and contrastive learning to close $N \longleftrightarrow D(H)$ and $N \longleftrightarrow R(N)$, respectively. Furthermore, the high-level adaptation stage adopts the same network architecture and detection loss function as the original YOLOX [5] and does not need ground-truth bounding boxes of hazy images through self-supervised learning.

2.2.1. Closing $N \longleftrightarrow D(H)$. Inspired by [50], we introduce an extra pretext task of solving jigsaw puzzles to understand the spatial context of objects. This widely used game can be simultaneously optimized with object classification across various source domains, thereby enhancing generalization through a straightforward multitask process.

As shown in Figure 2, we first decompose each dehazed image $D(H)$ and clean image N into 3×3 patches and set the patch permutation number to 30. Subsequently, we can represent the loss function for closing $N \longleftrightarrow D(H)$ as follows:

$$L_{N \longleftrightarrow D(H)} = L_c\left(F_{\text{jig}}^{D(H)}, p_{\text{jig}}^{D(H)}\right) + L_c\left(F_{\text{jig}}^{N}, p_{\text{jig}}^{N}\right), \tag{8}$$

where p_{jig} and F_{jig} are permutation labels and features extracted from the corresponding domain, respectively. Meanwhile, L_c denotes the cross-entropy loss. Notably, $D(H)$ and N utilize the same jigsaw classification head to enforce the mapping of semantic features into the same space, thereby closing the high-level gaps.

2.2.2. Closing $N \longleftrightarrow R(N)$. Since $R(N)$ is generated from N, we make full use of the inherent information of the images to bring them closer via the contrastive learning classifier [51]. Especially, given a query v with "positive" pair v^+ and "negative" pairs $v^- = \{v_1^-, v_2^-, \ldots, v_N^-\}$, the similarity objective $L_p(v, v^+, v^-)$ can be measured by dot product as follows:

$$\begin{aligned} L_p(v, v^+, v^-) &= -\log\left[\frac{\sigma(v, v^+)}{\sigma(v, v^+) + \sum_{n=1}^{N}\sigma(v, v_n^-)}\right], \\ \sigma(x, y) &= \exp\left(\frac{x \cdot y}{\tau}\right), \end{aligned} \tag{9}$$

where τ denotes a temperature hyperparameter. To close $N \longleftrightarrow R(N)$, we select the positive pair of N as the patch from $R(N)$ and vice versa as follows:

$$L_{N \longleftrightarrow R(N)} = L_p(N, R(N)^+, N^-) + L_p(R(N), N^+, R(N)^-). \tag{10}$$

2.2.3. Loss Function. We adopt a multitask loss function to constrain the self-supervised training process, which is defined as follows:

$$L_{\text{high-level}} = L_{\text{det}} + \lambda_{\text{jig}} L_{N \longleftrightarrow D(H)} + \lambda_{\text{con}} L_{N \longleftrightarrow R(N)}, \tag{11}$$

where λ_{jig} and λ_{con} are balanced weights that we experimentally set to 0.5 and 0.5, respectively. L_{det} denotes the detection loss, which is the same as the original YOLOX [5].

3. Experiments

3.1. Experimental Settings

3.1.1. Dataset. We conduct experiments on two synthetic foggy datasets (VOC_Foggy [28] and Foggy Cityscapes [60]), as well as two real-world foggy datasets (Foggy Driving dataset [60] and RTTS [61]). The VOC_Foggy dataset [28] is proposed based on the classic VOC dataset [62] according to the atmospheric scattering model [43]. Followed by [28], we filter out images containing five classes of objects from VOC2007_trainval and VOC2012_trainval, to build VOC_norm_trainval. The VOC_norm_test is selected from the VOC2007_test in a similar way. The Foggy Driving dataset [60] is a real-world foggy dataset involving 466 vehicle instances and 269 human instances that are labeled from 101 real-world foggy images. The RTTS dataset [61] collects 4322 real-world hazy images with five annotated object classes. The RTTS dataset is commonly used to evaluate the performance of dehazing algorithms in real-world scenarios from a task-driven perspective. Similar to [28, 31, 63], real-world foggy datasets are only used for testing object detection. To ensure consistency between training and testing, we adopt the same protocol as [28] and evaluate the performance on the aforementioned three datasets based on the five object classes of person, bicycle, car, motorbike, and bus. Foggy Cityscapes dataset [60] is created by applying fog synthesis on the Cityscapes dataset [64]. It has the same data split as the Cityscapes dataset [64], i.e., 2975 images for training and 500 images for testing, and all 8 categories are considered.

3.1.2. Training Details. In the low-level adaptation stage, we employ the SGD optimizer with a weight decay of $5e^{-4}$ and a momentum of 0.9. The network is optimized with a learning rate of $1e^{-4}$ and a batch size of 4. In the high-level adaptation stage, we still use a similar SGD optimizer while the learning rate is set to $5e^{-4}$. Note that our model is allowed to use the labels of normal images but is not allowed to use the labels of hazy images. Additionally, our memory bank is represented by a queue with the size of 128.

3.1.3. Evaluation Settings. We compare our HLA-HOD with eighteen state-of-the-art object detection approaches, which can be classified into three categories: (i) baseline YOLOX [5] and FRCNN [10]; (ii) low-level adaptation methods: DCP [16] + YOLOX, AOD [21] + YOLOX [5], Semi [52] + YOLOX [5], FFA [53] + YOLOX [5], MSB [30] + YOLOX [5], AEC [54] + YOLOX [5], DCP [16] + FRCNN [10], FFA [53] + FRCNN [10], MSB [30] + FRCNN [10], IA-YOLO [28], and DS-Net [31]; (iii) high-level adaptation methods: OSHOT [55], GPA [36], UMT [56], SAPNet [57], and EPM [58]. We note that HLA-HOD and all high-level adaptation methods do not use the labels of hazy images.

3.2. Comparisons

3.2.1. Comparison on the Synthetic Dataset. We first compare our HLA-HOD with eighteen detection algorithms on

Table 1: Comparison of HLA-HOD with state-of-the-art detection models on the VOC-FOG-test dataset.

Methods	Type	Backbone	Person	Bicycle	Car	Motorbike	Bus	mAP
YOLOX [5]	Baseline	—	79.97	67.95	74.75	58.62	83.12	72.88
FRCNN [10]	Baseline	—	75.39	71.91	70.12	60.58	81.27	71.85
DCP [16] + YOLOX [5]	Low-level	YOLOX [5]	*81.58*	78.80	79.75	78.51	*85.64*	*80.86*
AOD [21] + YOLOX [5]	Low-level	YOLOX [5]	81.26	73.56	76.98	71.18	83.08	77.21
Semi [52] + YOLOX [5]	Low-level	YOLOX [5]	81.15	76.94	76.92	72.89	84.88	78.56
FFA [53] + YOLOX [5]	Low-level	YOLOX [5]	78.30	70.31	69.97	68.80	80.72	73.62
MSB [30] + YOLOX [5]	Low-level	YOLOX [5]	74.91	72.89	70.73	**80.92**	79.81	75.85
AEC [54] + YOLOX [5]	Low-level	YOLOX [5]	79.23	70.57	73.14	72.05	80.46	75.09
DCP [16] + FRCNN [10]	Low-level	FRCNN [10]	77.34	*78.99*	74.36	79.33	83.20	78.64
FFA [53] + FRCNN [10]	Low-level	FRCNN [10]	75.06	71.02	68.49	70.28	79.95	72.96
MSB [30] + FRCNN [10]	Low-level	FRCNN [10]	73.67	74.80	69.76	78.96	80.11	75.46
IA-YOLO [28]	Low-level	YOLOX [5]	70.98	61.98	70.98	57.93	61.98	64.77
DS-Net [31]	Low-level	RetinaNet [15]	72.44	60.47	*81.27*	53.85	61.43	65.89
OSHOT [55]	High-level	FRCNN [10]	70.42	70.52	73.84	67.33	66.99	69.82
GPA [36]	High-level	FRCNN [10]	60.37	36.80	39.31	34.48	33.31	40.85
UMT [56]	High-level	FRCNN [10]	75.22	72.98	76.66	76.95	76.17	75.60
SAPNet [57]	High-level	FRCNN [10]	62.74	62.25	70.06	59.81	60.28	63.03
EPM [58]	High-level	FCOS [59]	66.74	56.22	72.70	44.79	61.88	60.47
HLA-HOD	High-low	YOLOX [5]	**83.20**	**80.12**	**84.78**	*80.10*	**88.96**	**83.43**

Bold and italic values are used to indicate the 1st and 2nd ranks, respectively.

the synthetic VOC_Foggy dataset for quantitative evaluation, as shown in Table 1. For fair comparison, we retrain all compared methods on the VOC_Foggy [60] dataset. For low-level adaptation approaches, it is observed that HLA-HOD outperforms all other state-of-the-art models even without ground-truth bounding boxes and clean images. Especially, HLA-HOD achieves a detection accuracy up to 2.57% mAP higher than that achieved by DCP + YOLOX. Compared with the baseline, some approaches even worsen the object detection performance, such as IA-YOLO and DS-Net. HLA-HOD improves the YOLOX baseline by 10.55% mAP on the VOC-FOG-test dataset. For high-level adaptation approaches, they also do not require the labels of hazy images. It is observed that our HLA-HOD still outperforms other approaches. This is because closing $N \longleftrightarrow D(H)$ and $N \longleftrightarrow R(N)$ is more effective than directly bridging the gap between normal and hazy images. Furthermore, we retrain all compared methods on the Foggy Cityscapes [60] dataset and conduct experiments on the testing set. The mAP threshold for all the models is set to 0.5. As shown in Table 2, HLA-HOD still outperforms all the competing methods by a large margin. Especially, HLA-HOD is better than the second best method UMT [56] by 4.76% mAP.

3.2.2. Comparison on the Real-World Dataset. We also compare our HLA-HOD with eighteen detection algorithms on two real-world foggy datasets, i.e., the Foggy Driving dataset and RTTS dataset. As demonstrated in Tables 3 and 4, our HLA-HOD still achieves the highest mAP results. Especially, HLA-HOD improves the baseline YOLOX by 3.36% and 3.02% mAP on the Foggy Driving dataset and RTTS dataset, respectively. The qualitative detection results are shown in Figure 4, and our HLA-HOD detects more objects with higher confidence, which demonstrates that our approach effectively handles both synthetic and real-world foggy scenarios.

3.3. Ablation Study. To verify the effectiveness of each module in our HLA-HOD, we conduct ablation experiments with different settings on three testing datasets as shown in Table 5.

3.3.1. Effect of Low-Level Adaptation. We first compare our low-level adaptation stage (LAS) with CycleGAN [44]. As depicted in Table 5, CycleGAN does not need ground-truth clean images, but its performance worsens in comparison to the baseline YOLOX [5] owing to color and structural distortions. Benefiting from our weather-specific decomposition strategy, LAS significantly enhances object detection performance under hazy conditions. Additionally, we also try different generators in LAS and observe that our modified RefineNet with the memory bank achieves the best results. Especially, when we remove the memory bank and replace the randomly generated β_N by a fixed value while keeping the other settings (denoted Refine$*$), the mAP results decrease from 83.43%, 36.95%, and 56.92% to 81.33%, 32.56%, and 52.89% on three testing datasets, respectively. The results clearly indicate that the utilization of random scattering coefficients significantly improves the generalization ability of the model across diverse haze densities. Additionally, we conduct ablation experiments on different loss functions on three testing datasets as shown in Table 6. The full loss function of HLA-HOD achieves the highest performance, indicating that all loss functions are beneficial for object detection under hazy conditions.

3.3.2. Effect of High-Level Adaptation. We further explore our strategies of closing $D(H) \longleftrightarrow N$ and $R(N) \longleftrightarrow N$ as shown in Table 5. Especially, we remove one component along with its corresponding loss function to each

Table 2: Comparison of HLA-HOD with state-of-the-art detection models on the Foggy Cityscapes dataset.

Methods	Person	Rider	Car	Truck	Bus	Train	Motor	Bicycle	mAP
YOLOX [5]	31.59	38.27	44.09	13.82	29.96	8.81	16.99	29.08	26.58
FRCNN [10]	30.16	38.79	36.20	19.33	31.98	9.25	23.27	32.84	27.73
DCP [16] + YOLOX [5]	39.77	42.05	50.35	24.39	38.70	35.61	30.76	28.68	36.29
AOD [21] + YOLOX [5]	37.63	40.27	51.67	22.14	38.28	32.85	31.39	29.90	35.52
Semi [52] + YOLOX [5]	38.29	40.11	51.24	23.73	39.14	34.52	30.35	32.25	36.20
FFA [53] + YOLOX [5]	32.04	38.39	43.58	17.92	36.20	28.53	28.02	26.50	31.40
MSB [30] + YOLOX [5]	37.48	*46.91*	54.69	27.65	49.96	42.74	32.34	35.06	40.85
AEC [54] + YOLOX [5]	38.82	44.38	52.50	23.38	48.39	41.12	31.85	38.57	39.88
DCP [16] + FRCNN [10]	38.56	43.49	48.06	16.49	49.18	37.44	33.56	29.21	37.00
FFA [53] + FRCNN [10]	29.30	42.74	37.57	14.76	36.59	27.29	30.58	27.86	30.84
MSB [30] + FRCNN [10]	34.92	46.70	49.56	24.10	50.06	45.42	*35.75*	*40.81*	40.92
IA-YOLO [28]	32.29	43.75	53.84	22.64	48.95	42.28	34.39	35.96	39.26
DS-Net [31]	37.91	42.65	50.46	18.32	38.79	33.58	31.29	32.95	35.74
OSHOT [55]	32.12	46.06	43.14	20.37	39.81	15.85	27.08	32.41	32.11
GPA [36]	32.87	46.69	54.12	24.65	45.72	41.13	32.39	38.65	39.53
UMT [56]	32.96	46.72	48.56	**34.12**	*56.47*	*46.82*	30.36	37.34	*41.67*
SAPNet [57]	40.78	46.66	*59.79*	24.31	46.82	37.47	30.43	40.69	40.87
EPM [58]	*41.92*	38.65	56.71	22.64	41.53	26.76	24.64	35.48	36.04
HLA-HOD	**43.90**	**49.08**	**60.84**	*33.52*	**56.97**	**48.14**	**36.73**	**42.25**	**46.43**

Table 3: Comparison of HLA-HOD with state-of-the-art detection models on the Foggy Driving dataset.

Methods	Person	Bicycle	Car	Motorbike	Bus	mAP
YOLOX [5]	26.69	23.04	56.27	2.38	41.98	30.07
FRCNN [10]	24.88	23.58	55.92	4.99	40.47	29.97
DCP [16] + YOLOX [5]	22.24	10.78	56.34	**7.14**	*50.66*	29.43
AOD [21] + YOLOX [5]	24.54	*33.82*	56.75	4.76	36.04	31.18
Semi [52] + YOLOX [5]	22.39	27.73	56.47	4.76	44.93	31.26
FFA [53] + YOLOX [5]	19.18	18.07	50.83	2.38	42.77	26.65
MSB [30] + YOLOX [5]	23.14	30.31	57.15	6.04	45.82	*32.49*
AEC [54] + YOLOX [5]	22.83	24.37	50.51	**7.14**	41.45	29.26
DCP [16] + FRCNN [10]	20.77	11.90	54.38	6.06	47.69	28.16
FFA [53] + FRCNN [10]	20.36	24.67	55.90	3.05	40.70	28.94
MSB [30] + FRCNN [10]	23.09	26.34	57.15	5.81	43.76	31.23
IA-YOLO [28]	16.20	11.76	41.43	4.76	17.55	18.34
DS-Net [31]	*26.74*	20.54	*58.16*	**7.14**	36.11	29.74
OSHOT [55]	20.46	18.18	40.84	3.03	34.15	23.33
GPA [36]	15.92	19.91	36.74	3.03	19.67	19.05
UMT [56]	19.71	20.98	44.60	0.47	39.13	24.98
SAPNet [57]	23.67	27.27	48.20	3.64	30.70	26.70
EPM [58]	26.50	21.21	43.75	0.21	27.01	23.74
HLA-HOD	**29.16**	**34.96**	**60.87**	*6.77*	**52.99**	**36.95**

configuration at one time. If we only use contrastive learning $R(N)\longleftrightarrow N$ or multitask learning $D(H)\longleftrightarrow N$, the performance drops. This is due to the fact that if we only focus on closing $D(H)\longleftrightarrow N$, the distance between $R(N)$ and N will increase and vice versa. These results also support our design of the high-level adaptation stage and the importance of establishing the intermediate domains $D(H)$ and $R(N)$.

3.4. Comparisons with CycleGAN. We compare the qualitative dehazed results of our low-level adaptation stage (LAS) with CycleGAN [44] as shown in Figure5. It can be observed that CycleGAN leads to color and structure distortions. If we replace our LAS with the CycleGAN structure, the gap between dehazed images and normal images will increase, thus leading to inferior object detection performance. It demonstrates the superiority of our LAS with

Table 4: Comparison of HLA-HOD with state-of-the-art detection models on the RTTS dataset.

Methods	Person	Bicycle	Car	Motorbike	Bus	mAP
YOLOX [5]	76.07	48.47	63.88	41.03	22.76	50.44
FRCNN [10]	72.69	49.53	61.44	40.19	22.68	49.31
DCP [16] + YOLOX [5]	*76.81*	50.03	62.84	40.62	23.73	50.81
AOD [21] + YOLOX [5]	76.49	43.32	61.03	34.54	22.16	47.51
Semi [52] + YOLOX [5]	75.71	46.72	62.74	40.37	24.51	50.01
FFA [53] + YOLOX [5]	76.52	48.13	64.31	39.74	23.71	50.48
MSB [30] + YOLOX [5]	73.97	50.91	*68.43*	44.32	26.78	*52.88*
AEC [54] + YOLOX [5]	75.42	49.26	65.81	38.69	26.78	51.19
DCP [16] + FRCNN [10]	73.26	51.23	59.82	38.89	24.21	49.48
FFA [53] + FRCNN [10]	72.10	49.88	60.57	39.33	21.70	48.72
MSB [30] + FRCNN [10]	73.48	50.67	61.48	40.79	23.01	49.89
IA-YOLO [28]	67.25	35.28	41.14	20.97	13.64	35.66
DS-Net [31]	68.81	18.02	46.13	15.15	15.44	32.71
OSHOT [55]	69.70	*54.30*	49.24	41.08	25.94	48.05
GPA [36]	55.16	41.82	39.05	31.65	20.78	37.69
UMT [56]	70.45	52.22	49.57	*50.65*	26.83	49.94
SAPNet [57]	69.80	50.22	51.98	44.91	*26.87*	48.76
EPM [58]	71.95	33.84	55.17	22.07	19.91	40.59
HLA-HOD	**80.29**	**56.22**	**70.01**	**51.02**	**27.06**	**56.92**

(a) (b) (c) (d) (e)

Figure 4: Qualitative comparison of HLA-HOD with state-of-the-art detection models on the real-world dataset. (a) YOLOX. (b) FFA+YOLOX. (c) IA-YOLO. (d) OSHOT. (e) HLA-HOD.

a decomposition strategy based on physical factors such as atmospheric light, transmission map, and scattering coefficient.

3.5. Analysis and Discussion. In this section, we provide the in-depth analysis and discussion about HLA-HOD. We first conduct experiments on different balanced weight configurations. Subsequently, we incorporate several commonly employed low-level and high-level techniques into our framework to verify the superiority of our high-level and low-level adaptation stages. Finally, we compare the running time of our HLA-HOD with that of other methods to demonstrate its efficiency.

3.5.1. The Effect of Different Balanced Weights. We evaluate the effectiveness of different balanced weights on the VOC_Foggy dataset. For fair comparison, the same training settings are kept for all models' testing. As shown in Figure 6, the balanced weights between different losses are chosen experimentally, which optimizes our network effectively.

3.5.2. The Effect of Different Low-Level Adaptations. In addition to CycleGAN, we explore the influence of different low-level adaptation techniques on the dehazed results, as presented in Table 7. For fair comparison, we incorporate these unsupervised approaches into the low-level adaptation stage and retrain the entire paradigm. Despite requiring no

Table 5: Performance comparison of different settings under foggy conditions.

Low-level		High-level		mAP		
CycleGAN	Ours	$D(H)\longleftrightarrow N$	$R(N)\longleftrightarrow N$	VOC-F	FD	RTTS
w/o	w/o	w/o	w/o	72.88	30.07	50.44
U-Net	w/o	✓	✓	68.38	28.44	49.76
Refine	w/o	✓	✓	69.10	30.22	51.08
Refine*	w/o	✓	✓	69.88	31.63	51.74
w/o	U-Net	✓	✓	78.95	31.17	51.82
w/o	Refine	✓	✓	80.06	31.45	52.71
w/o	Refine*	✓	✓	81.33	32.56	52.89
w/o	Refine(M)	✓	w/o	81.36	31.18	51.03
w/o	Refine(M)	w/o	✓	80.02	30.54	51.44
w/o	Refine(M)	✓	✓	83.43	36.95	56.92

Note that the third row represents the YOLOX backbone [5], Refine, Refine*, and Refine(M) denote the original RefineNet, modified RefineNet with self-attention layers, and modified RefineNet with the memory bank, respectively. w/o represents without the model.

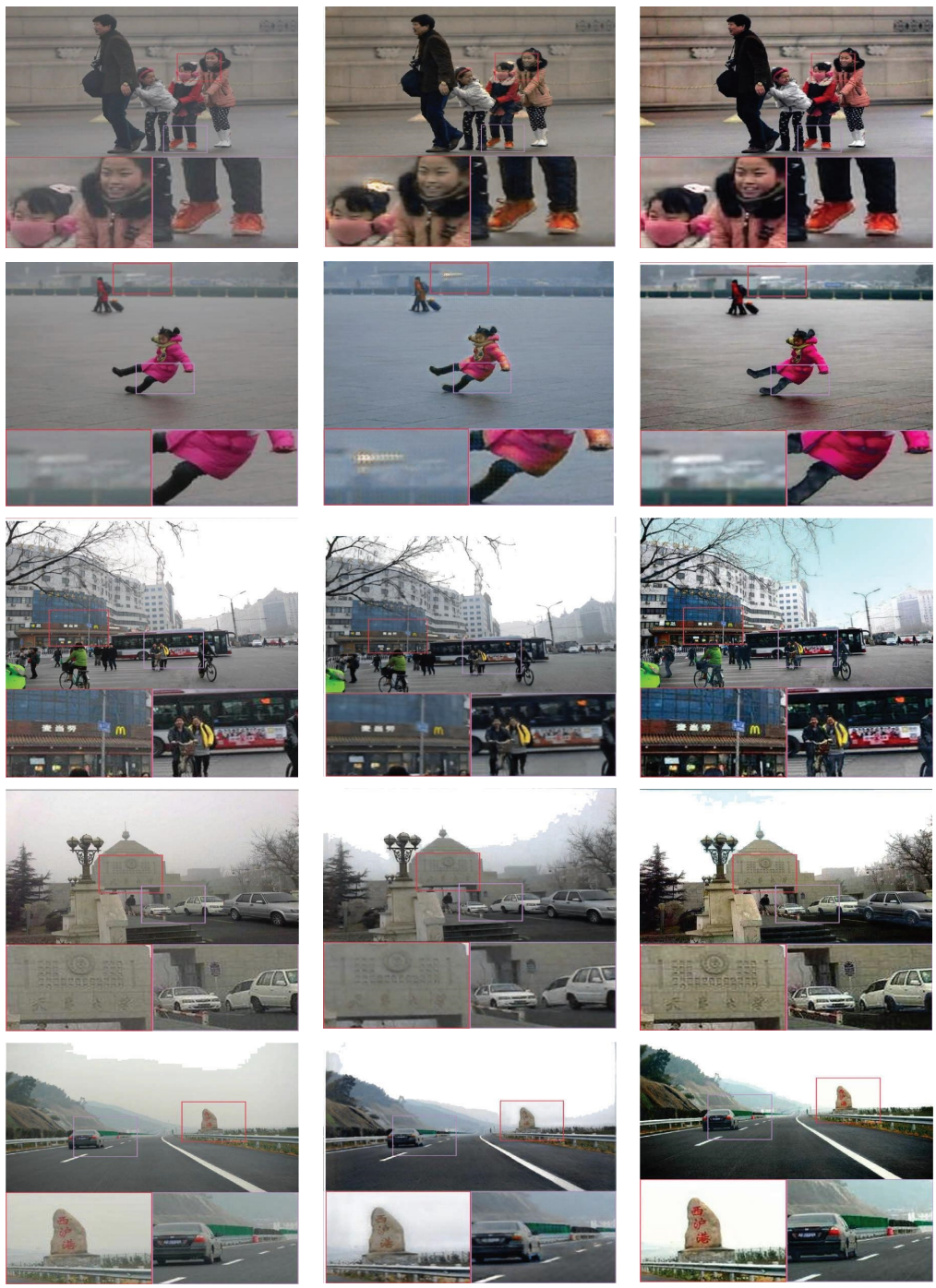
Table 6: Performance comparison of different loss functions under foggy conditions.

Variants	V_0	V_1	V_2	V_3
L_{cyc}	✓	✓	✓	✓
L_{adv}	w/o	✓	✓	✓
L_{dep}	w/o	w/o	✓	✓
L_{sca}	w/o	w/o	w/o	✓
VOC-F	79.62	80.06	81.34	83.43
FD	29.99	30.48	31.72	36.95
RTTS	50.37	51.05	52.68	56.92

ground-truth clean images, the performance of these models significantly drops, further verifying the superiority of our low-level adaptation method.

3.5.3. The Effect of Different High-Level Adaptations. We demonstrate the effect of different high-level adaptations for closing $R(N)\longleftrightarrow N$ and $D(H)\longleftrightarrow N$, as presented in Table 8. We make the following observations: (i) the use of jigsaw for $D(H)\longleftrightarrow N$ and contrastive learning for $R(N)\longleftrightarrow N$ exhibits superior performance. (ii) Using other high-level adaptations approaches, such as rotation [68] and pseudolabels [40], can damage the performance. (iii) While using jigsaw for $R(N)\longleftrightarrow N$ and contrastive learning for $D(H)\longleftrightarrow N$ yields the second best performance, the level of improvement remains inferior to that of our method.

3.5.4. Running Time Comparison. We compare the running time of our method with different approaches presented in Table 9. We evaluate the models using 100 images of size 512×512. The proposed method outperforms most of the combination models of object detection and dehazing by a significant margin. Specifically, HLA-HOD runs an average of 106 ms, which is faster than DCP + YOLOX, semi + YOLOX, FFA + YOLOX, MSB + YOLOX, and



(a) (b) (c)

Figure 5: Qualitative comparisons with CycleGAN on the real-world dataset. (a) Input. (b) CycleGAN. (c) Ours.

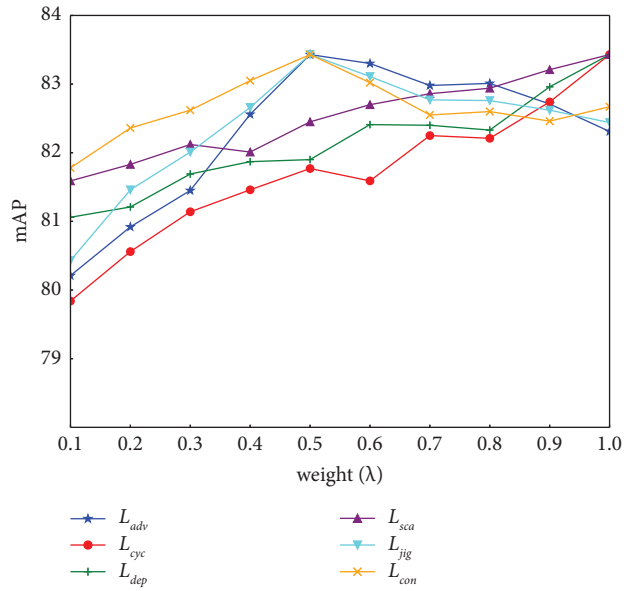


Figure 6: Effectiveness of different balanced weights.

Table 7: Performance comparison of different low-level adaptation methods under foggy conditions.

Low-level	mAP		
	VOC-F	FD	RTTS
Cycle-dehaze [23]	79.49	29.73	49.91
Deep DCP [65]	80.54	32.48	51.68
YOLY [66]	79.57	30.78	50.39
PSD [67]	80.36	31.82	51.14
Ours	83.43	36.95	56.92

Table 8: Performance comparison of different high-level adaptation methods under foggy conditions.

High-level		mAP		
$D(H)\longleftrightarrow N$	$R(N)\longleftrightarrow N$	VOC-F	FD	RTTS
Rotation [68]	Contrastive	78.19	28.63	48.74
Pseudolabels [40]	Contrastive	79.20	31.06	51.35
Jigsaw	Rotation [68]	79.82	29.59	50.45
Jigsaw	Pseudolabels [40]	80.13	31.36	51.32
Contrastive	Jigsaw	81.51	34.38	54.64
Jigsaw	Contrastive	83.43	36.95	56.92

Table 9: Comparisons of the running times (seconds).

Methods	Time (s)
YOLOX	0.042
DCP + YO	0.110
AOD + YO	0.072
Semi + YO	0.137
FFA + YO	0.119
MSB + YO	0.147
AEC + YO	0.132
IA-YOLO	0.096
OSHOT	0.152
GPA	0.209
UMT	0.172
SAPNet	0.092
EPM	0.120
Ours	0.106

AEC + YOLOX by 4 ms, 31 ms, 13 ms, 41 ms, and 26 ms, respectively. Although our approach may not be the fastest among the high-level adaptation methods, its performance is still acceptable. In summary, although our HLA-HOD is composed of high-level and low-level adaptations, it provides the best trade-off between inference time and object detection accuracy compared to other methods.

4. Conclusion

In this paper, we propose a Joint High-Low Adaptation Object Detection paradigm (HLA-HOD) under hazy weather conditions. Unlike other low-level adaptation and high-level adaptation approaches, HLA-HOD combines both low-level adaptation and high-level adaptation to achieve superior performance on hazy images even without ground-truth bounding boxes and clean images. In the low-level adaptation stage, we leverage the physical properties of the real-world hazy environment to generate the dehazed and rehazed results, which are used to set up two intermediate states for high-level adaptation. In the high-level adaptation stage, we push the feature spaces of dehazed, rehazed, and normal states towards each other, instead of directly closing the gap between normal and hazy states. Extensive experiments prove the potential of joint high-low adaptation and the superiority of HLA-HOD.

Data Availability

The data used to support the findings of this study are available at VOC_Foggy (https://github.com/wenyyu/Image-Adaptive-YOLO), Foggy Cityscapes (https://people.ee.ethz.ch/~csakarid/SFSU_synthetic/), Foggy Driving dataset (https://people.ee.ethz.ch/~csakarid/SFSU_synthetic/), and RTTS (https://sites.google.com/view/reside-dehaze-datasets/reside-v0).

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China (no. 2020YFB1805400), in part by the National Natural Science Foundation of China (no. 42071431), and in part by the Provincial Key Research and Development Program of Hubei, China (no. 2020BAB101).

References

[1] I. Bozcan and E. Kayacan, "Context-dependent anomaly detection for low altitude traffic surveillance," in *Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 224–230, Xi'an, China, June 2021.

[2] C. Chen, A. Seff, A. Kornhauser, and J. Xiao, "Deepdriving: learning affordance for direct perception in autonomous driving," in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 2722–2730, Santiago, Chile, December 2015.

[3] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," pp. 779–788, 2016, https://arxiv.org/abs/1506.02640.

[4] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," pp. 7263–7271, 2017, https://arxiv.org/abs/1612.08242.

[5] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "Yolox: Exceeding yolo Series in 2021," 2021, https://arxiv.org/abs/2107.08430.

[6] D. Kim, S. Kim, S. Jeong et al., "Rotational multipyramid network with bounding-box transformation for object detection," *International Journal of Intelligent Systems*, vol. 36, no. 9, pp. 5307–5338, 2021.

[7] N. Wen, R. Guo, D. Ma, X. Ye, and B. He, "AIoU: adaptive bounding box regression for accurate oriented object detection," *International Journal of Intelligent Systems*, vol. 37, no. 1, pp. 748–769, 2022.

[8] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," pp. 580–587, 2014, https://arxiv.org/abs/1311.2524.

[9] R. Girshick, "Fast r-cnn," in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1440–1448, Santiago, Chile, December 2015.

[10] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[11] W. Liu, D. Anguelov, and D. Erhan, "Ssd: single shot multibox detector," pp. 21–37, 2016, https://arxiv.org/abs/1512.02325.

[12] W. Wang, W. Yang, and J. Liu, "Hla-face: joint high-low adaptation for low light face detection," pp. 16195–16204, 2021, https://arxiv.org/abs/2104.01984.

[13] Z. Shen, Z. Liu, J. Li, Y. G. Jiang, Y. Chen, and X. Xue, "Dsod: learning deeply supervised object detectors from scratch," pp. 1919–1927, 2017, https://arxiv.org/abs/1708.01241.

[14] C. Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "Dssd: deconvolutional single shot detector," 2017, https://arxiv.org/abs/1701.06659.

[15] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," pp. 2980–2988, 2017, https://arxiv.org/abs/1708.02002.

[16] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 12, pp. 2341–2353, 2011.

[17] Q. Zhu, J. Mai, and L. Shao, "A fast single image haze removal algorithm using color attenuation prior," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 3522–3533, 2015.

[18] R. T. Tan, "Visibility in bad weather from a single image," in *Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, Anchorage, AK, USA, June 2008.

[19] B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao, "Dehazenet: an end-to-end system for single image haze removal," *IEEE Transactions on Image Processing*, vol. 25, no. 11, pp. 5187–5198, 2016.

[20] W. Ren, S. Liu, H. Zhang, J. Pan, X. Cao, and M. H. Yang, "Single image dehazing via multi-scale convolutional neural networks," in *Proceedings of the Computer Vision--ECCV 2016: 14th European Conference*, pp. 154–169, Amsterdam, The Netherlands, October 2016.

[21] B. Li, X. Peng, Z. Wang, J. Xu, and D. Feng, "Aod-net: all-in-one dehazing network," in *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 4770–4778, Venice, Italy, October 2017.

[22] F. Deeba, F. A. Dharejo, M. Zawish et al., "A novel image dehazing framework for robust vision-based intelligent systems," *International Journal of Intelligent Systems*, vol. 37, no. 12, pp. 10495–10513, 2022.

[23] D. Engin, A. Genç, and H. Kemal Ekenel, "Cycle-dehaze: enhanced cyclegan for single image dehazing," in *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 825–833, Salt Lake City, UT, USA, June 2018.

[24] W. Liu, X. Hou, J. Duan, and G. Qiu, "End-to-end single image fog removal using enhanced cycle consistent adversarial networks," *IEEE Transactions on Image Processing*, vol. 29, pp. 7819–7833, 2020.

[25] Y. Yang, C. Wang, R. Liu, L. Zhang, X. Guo, and D. Tao, "Self-augmented unpaired image dehazing via density and depth decomposition," in *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2037–2046, New Orleans, LA, USA, June 2022.

[26] J. M. J. Valanarasu, R. Yasarla, and V. M. Patel, "Transweather: Transformer-based restoration of images degraded by adverse weather conditions," pp. 2353–2363, 2022, https://arxiv.org/abs/2111.14813.

[27] Y. Song, Z. He, H. Qian, and X. Du, "Vision transformers for single image dehazing," 2022, https://arxiv.org/abs/2204.03883.

[28] W. Liu, G. Ren, R. Yu, S. Guo, J. Zhu, and L. Zhang, "Image-adaptive yolo for object detection in adverse weather conditions," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 2, pp. 1792–1800, 2022.

[29] Y. Lee, J. Jeon, Y. Ko, B. Jeon, and M. Jeon, "Task-driven deep image enhancement network for autonomous driving in bad weather," pp. 13746–13753, 2021, https://arxiv.org/abs/2110.07206.

[30] H. Dong, J. Pan, and L. Xiang, "Multi-scale boosted dehazing network with dense feature fusion," in *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2157–2167, Seattle, WA, USA, June 2020.

[31] S. C. Huang, T. H. Le, and D. W. Jaw, "DSNet: joint semantic learning for object detection in inclement weather conditions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 8, pp. 2623–2633, 2021.

[32] S. C. Huang, Q. V. Hoang, and T. H. Le, "SFA-Net: a selective features absorption network for object detection in rainy weather conditions," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–11, 2022.

[33] M. Hnewa and H. Radha, "Multiscale domain adaptive yolo for cross-domain object detection," pp. 3323–3327, 2021, https://arxiv.org/abs/2106.01483.

[34] F. Rezaeianaran, R. Shetty, R. Aljundi, D. O. Reino, S. Zhang, and B. Schiele, "Seeking similarities over differences: similarity-based domain alignment for adaptive object detection," pp. 9204–9213, 2021, https://arxiv.org/abs/2110.01428.

[35] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. Van Gool, "Domain adaptive faster r-cnn for object detection in the wild," pp. 3339–3348, 2018, https://arxiv.org/abs/1803.03243.

[36] M. Xu, H. Wang, B. Ni, Q. Tian, and W. Zhang, "Cross-domain detection via graph-induced prototype alignment," pp. 12355–12364, 2020, https://arxiv.org/abs/2003.12849.

[37] Z. He and L. Zhang, "Multi-adversarialfaster-rcnn for unrestricted object detection," pp. 6668–6677, 2019, https://arxiv.org/abs/1907.10343.

[38] K. Saito, Y. Ushiku, T. Harada, and K. Saenko, "Strong-weak distribution alignment for adaptive object detection," pp. 6956–6965, 2019, https://arxiv.org/abs/1812.04798.

[39] R. Yu, W. Liu, Y. Zhang, Z. Qu, D. Zhao, and B. Zhang, "Deepexposure: learning to expose photos with asynchronously reinforced adversarial learning," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[40] N. Inoue, R. Furuta, T. Yamasaki, and K. Aizawa, "Cross-domainweakly-supervised object detection through progressive domain adaptation," pp. 5001–5009, 2018, https://arxiv.org/abs/1803.11365.

[41] M. Khodabandeh, A. Vahdat, M. Ranjbar, and W. G. Macready, "A robust learning approach to domain adaptive object detection," pp. 480–490, 2019, https://arxiv.org/abs/1904.02361.

[42] S. G. Narasimhan and S. K. Nayar, "Chromatic framework for vision in bad weather," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No.PR00662)*, pp. 598–605, Hilton Head, SC, USA, June 2000.

[43] S. G. Narasimhan and S. K. Nayar, "Vision and the atmosphere," *International Journal of Computer Vision*, vol. 48, no. 3, pp. 233–254, 2002.

[44] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," pp. 2223–2232, 2017, https://arxiv.org/abs/1703.10593.

[45] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: multi-path refinement networks for high-resolution semantic segmentation," pp. 1925–1934, 2017, https://arxiv.org/abs/1611.06612.

[46] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," pp. 770–778, 2016, https://arxiv.org/abs/1512.03385.

[47] A. Vaswani, N. Shazeer, and N. Parmar, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[48] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," pp. 2794–2802, 2017, https://arxiv.org/abs/1611.04076.

[49] Y. Wang, X. Yan, D. Guan et al., "Cycle-SNSPGAN: towards real-world image dehazing via cycle spectral normalized soft likelihood estimation patch GAN," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 11, pp. 20368–20382, 2022.

[50] F. M. Carlucci, A. D'Innocente, S. Bucci, B. Caputo, and T. Tommasi, "Domain generalization by solving jigsaw puzzles," pp. 2229–2238, 2019, https://arxiv.org/abs/1903.06864.

[51] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," pp. 9729–9738, 2020, https://arxiv.org/abs/1911.05722.

[52] L. Li, Y. Dong, W. Ren et al., "Semi-supervised image dehazing," *IEEE Transactions on Image Processing*, vol. 29, pp. 2766–2779, 2020.

[53] X. Qin, Z. Wang, Y. Bai, X. Xie, and H. Jia, "FFA-Net: feature fusion attention network for single image dehazing," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, pp. 11908–11915, 2020.

[54] H. Wu, Y. Qu, and S. Lin, "Contrastive learning for compact single image dehazing," pp. 10551–10560, 2021, https://arxiv.org/abs/2104.09367.

[55] A. D'Innocente, F. C. Borlino, S. Bucci, B. Caputo, and T. Tommasi, "One-shot unsupervised cross-domain detection," pp. 732–748, 2020, https://arxiv.org/abs/2005.11610.

[56] J. Deng, W. Li, Y. Chen, and L. Duan, "Unbiased mean teacher for cross-domain object detection," pp. 4091–4101, 2021, https://arxiv.org/abs/2003.00707.

[57] C. Li, D. Du, and L. Zhang, "Spatial attention pyramid network for unsupervised domain adaptation," pp. 481–497, 2020, https://arxiv.org/abs/2003.12979.

[58] C. C. Hsu, Y. H. Tsai, Y. Y. Lin, and M. H. Yang, "Every pixel matters: center-aware feature alignment for domain adaptive object detector," pp. 733–748, 2020, https://arxiv.org/abs/2008.08574.

[59] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: fully convolutional one-stage object detection," pp. 9627–9636, 2019, https://arxiv.org/abs/1904.01355.

[60] C. Sakaridis, D. Dai, and L. Van Gool, "Semantic foggy scene understanding with synthetic data," *International Journal of Computer Vision*, vol. 126, no. 9, pp. 973–992, 2018.

[61] B. Li, W. Ren, D. Fu et al., "Benchmarking single-image dehazing and beyond," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 492–505, 2019.

[62] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[63] V. A. Sindagi, P. Oza, R. Yasarla, and V. M. Patel, "Prior-based domain adaptive object detection for hazy and rainy conditions," pp. 763–780, 2020, https://arxiv.org/abs/1912.00070.

[64] M. Cordts, M. Omran, and S. Ramos, "The cityscapes dataset for semantic urban scene understanding," pp. 3213–3223, 2016, https://arxiv.org/abs/1604.01685.

[65] A. Golts, D. Freedman, and M. Elad, "Unsupervised single image dehazing using dark channel prior loss," *IEEE Transactions on Image Processing*, vol. 29, pp. 2692–2701, 2020.

[66] B. Li, Y. Gou, S. Gu, J. Z. Liu, J. T. Zhou, and X. Peng, "You only look yourself: unsupervised and untrained single image dehazing neural network," *International Journal of Computer Vision*, vol. 129, no. 5, pp. 1754–1767, 2021.

[67] Z. Chen, Y. Wang, Y. Yang, and D. Liu, "PSD: principled synthetic-to-real dehazing guided by physical priors," in *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7180–7189, Nashville, TN, USA, June 2021.

[68] N. Komodakis and S. Gidaris, "Unsupervised representation learning by predicting image rotations," 2018, https://arxiv.org/abs/1803.07728.