Detection-Friendly Dehazing: Object Detection in Real-World Hazy Scenes

Chengyang Li, *Graduate Student Member, IEEE*, Heng Zhou, *Graduate Student Member, IEEE*, Yang Liu, Caidong Yang, Yongqiang Xie, Zhongbo Li, and Liping Zhu

***Abstract*—Adverse weather conditions in real-world scenarios lead to performance degradation of deep learning-based detection models. A well-known method is to use image restoration methods to enhance degraded images before object detection. However, how to build a positive correlation between these two tasks is still technically challenging. The restoration labels are also unavailable in practice. To this end, taking the hazy scene as an example, we propose a union architecture BAD-Net that connects the dehazing module and detection module in an end-to-end manner. Specifically, we design a two-branch structure with an attention fusion module for fully combining hazy and dehazing features. This reduces bad impacts on the detection module when the dehazing module performs poorly. Besides, we introduce a self-supervised haze robust loss that enables the detection module to deal with different degrees of haze. Most importantly, an interval iterative data refinement training strategy is proposed to guide the dehazing module learning with weak supervision. BAD-Net improves further detection performance through detection-friendly dehazing. Extensive experiments on RTTS and VOChaze datasets show that BAD-Net achieves higher accuracy compared to the recent state-of-the-art methods. It is a robust detection framework for bridging the gap between low-level dehazing and high-level detection.**

***Index Terms*—Dehazing, detection-friendly, object detection, real world, weakly-supervised.**

I. Introduction

DEEP learning models tend to overfit the training data and have poor generalization ability. The designed vision models (*e.g.*, object detection, target tracking) are difficult to deal with complex and changeable real-world scenes [1]. The collected images are often affected by various environmental degradation factors, such as noise, blur, bad weather, and so on [2]. When the image quality is reduced, the subsequent feature extraction and analysis in high-level vision tasks are influenced. Therefore, taking a common hazy scene as an example, we attempt to explore how to improve detection accuracy under degraded conditions.

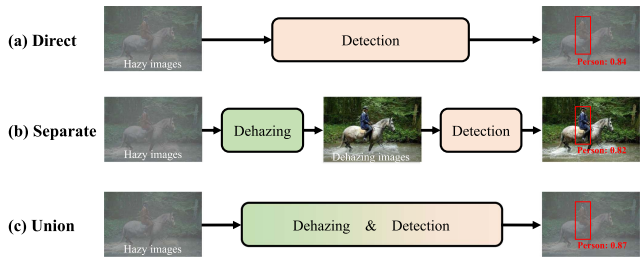

Fig. 1. Method comparison of object detection with dehazing in hazy conditions. (a) Detection models are directly trained on hazy images without restoration labels. (b) Dehazing models are first trained with restoration labels, and then enhanced dehazing images are input into pre-trained detection models. (c) Dehazing and detection models are jointly trained on hazy images with or without restoration labels.

With clear image labels, most approaches (Fig. 1(b)) first train the restoration models to generate images with no degradation. Then, the enhanced images are input into pre-trained detection models for bounding box regression. Few works [3], [4] optimize the entire model with joint loss of restoration and detection, as shown in Fig. 1(c). Although this strategy conforms to human logic, it has not achieved ideal detection results in contrastive experiments [5]. Since the purposes of restoration and detection are different, there are potential conflicts between them. Existing restoration models are usually based on neural networks. Their generated images may contain noise that is invisible to human eyes and harmful to subsequent detection models. This is similar to the principle of image adversarial attack [6]. Besides, restoration labels are often unavailable in real scenes. One simple resolution is to directly train detection models on degraded images, as shown in Fig. 1(a). This strategy often depends on the feature extraction ability of detection models. Recently, IA-YOLO [5] only use detection loss for training the union model of restoration and detection. However, it is difficult to meet the needs of universality and robustness. Therefore, how to design a union optimization model with high performance and fast convergence is an important research direction.

To this end, we propose a union network with two-branch attention (BAD-Net) that effectively combines the dehazing

Manuscript received 26 April 2022; revised 29 September 2022; accepted 3 January 2023. Date of publication 9 January 2023; date of current version 5 June 2023. Recommended for acceptance by V. Lempitsky. *(Corresponding authors: Yongqiang Xie; Zhongbo Li.)*

Chengyang Li is with the School of Computer Science, PKU, Beijing 100084, China, and also with the Institute of Systems Engineering, AMS, Beijing 100071, China (e-mail: chengyang_li@stu.pku.edu.cn).

Heng Zhou is with the School of Electronic Engineering, XDU, Xi'an, Shaanxi 710071, China, and also with the Institute of Systems Engineering, AMS, Beijing 100071, China (e-mail: hengzhou@stu.xidian.edu.cn).

Yang Liu, Caidong Yang, Yongqiang Xie, and Zhongbo Li are with the Institute of Systems Engineering, AMS, Beijing 100071, China (e-mail: jkwxsliuyang@outlook.com; yangcd2022@outlook.com; yqxie2021@outlook.com; zbli2021@outlook.com).

Liping Zhu is with the Beijing Key Laboratory of Petroleum Data Mining, CUP, Beijing 102249, China (e-mail: zhuliping@cup.edu.cn).

Digital Object Identifier 10.1109/TPAMI.2023.3234976

module and detection module. To make the entire model converge stably, we change the single-branch serial structure in previous methods to a two-branch parallel structure. Based on this, we design an attention fusion module to fully fuse dehazing and hazy features. It greatly enhances the complementarity of target features for subsequent object detection. Besides, a self-supervised haze robust loss is introduced to guide feature learning of the dehazing module. It constrains that target features of dehazing images extracted from the detection backbone are similar to those of hazy images. Thus, it enhances the robustness of the detection module. Moreover, we propose an interval iterative training strategy for further guiding dehazing module learning. The training input data are changed between the original and refined datasets at one-epoch intervals. The refined hazy dataset is formed by adding haze on dehazing images, which is output by the dehazing module trained at the previous epoch.

In summary, there are three main contributions:

- We explore how to build a positive correlation between low-level dehazing and high-level detection without clear image labels. This is essential yet under-studied in this field.
- We propose a bilinear union architecture, *i.e.*, BAD-Net, and an interval iterative training strategy. Both of them achieve detection-friendly dehazing, resulting in further detection performance improvement.
- In experiments, we are the first to divide and compare union methods from the perspective of training loss and strategy. Comparison results show that BAD-Net outperforms state-of-the-art methods on RTTS and VOChaze datasets.

The rest paper is structured as follows. Section II presents previous works of object detection, dehazing, attention mechanisms, and the relation between restoration and recognition. Section III introduces the details of proposed BAD-Net, while Section IV presents the experimental results on two datasets. In Section V, a conclusion is obtained.

II. Related Work

A. Recognition With Restoration

In the degraded environment, it's known that recognition may get better performance on the enhanced image through restoration (such as denoising, deblurring, deraining, dehazing, and so on). However, most degraded datasets are only used to evaluate image quality assessment (*e.g.*, PSNR, SSIM) for image restoration [7]. Few works pay attention to the improvement of recognition performance after restoration. In this work, we offer a new insight to investigate the logical relationship between image restoration and object recognition. At present, relevant works are divided into three main directions:

1) Recognition algorithms are directly trained on degraded images [8], [9], [10], [11].
2) Restoration algorithms are first trained to enhance degraded images, and then pre-trained recognition algorithms are evaluated on enhanced images [12], [13], [14], [15], [16], [17].
3) Restoration and recognition algorithms are jointly trained and optimized on degraded images [3], [4], [5], [18], [19], [20].

In the first direction [8], [9], [10], [11], recognition models are directly trained and evaluated on degraded datasets. Due to blurring or occlusion of targets under degradation, models learn fewer useful features for recognition. The difficulty of learning makes models difficult to converge during training. Therefore, the final recognition performance is poor and not easy to extend or promote. [8] proposes a dual-channel convolutional architecture to deal with quality degradations. [9] designs novel datasets "ImageNet-C" and "ImageNet-P," which are used to benchmark model's robustness to common perturbations. [10] proposes a dual directed capsule network for image recognition with low resolution. [11] designs an FPN-based architecture with a progressive top-down interaction and attention refinement module. It improves the robustness in non-optimal weather conditions.

For the second direction [12], [13], [14], [15], [16], [17], there are plenty of studies. Restoration and recognition are separated so that it is regarded as a two-stage strategy. It is difficult for experiments to indicate the relation and influence between them. Most works only adopt the performance of object recognition as a quantitative evaluation metric of restoration algorithms. These two-stage methods may improve recognition to a certain extent. However, they have low robustness and generalization. [12] explores the usefulness of super-resolution for other vision applications, such as edge detection, segmentation, and so on. Perceptual metrics and recognition performance are interrelated, but they are not fully substituted for each other. [13], [15] and [16] use recognition to evaluate dehazing for hazy images. [15] indicates that most dehazing methods sometimes may reduce recognition performance. Different from these methods, we aim at improving the performance of final detection rather than image restoration.

Few works concentrate on the third direction [3], [4], [5], [18], [19], [20]. The improvement of recognition performance is the largest among these three directions. Our work attempts to explore a connection between image restoration (low-level) and recognition (high-level). [3] designs a cascaded network combining denoising and high-level tasks. It adopts a joint loss for only updating the denoising module during training. [4] designs a dual-subnet network, which contains a detection subnet and a restoration subnet. In particular, the restoration subnet shares feature extraction layers with the detection subnet for prediction. Different from above, [5] trains the whole network with only a detection loss. Similar to [5], we propose a detection method without restoration labels to achieve detection-friendly dehazing.

B. Object Detection

Object detection plays a vital role in scene recognition and modeling. It is used to detect and recognize targets in various complex scenes, such as people, vehicles, and other objects. Recently, due to deep learning, the performance of object detection

has been greatly improved. Thus, DNN-based object detectors are currently the most general solution. Object detection methods based on deep learning are mainly divided into five aspects, such as single-stage, two-stage, multi-stage, anchor-free, and Transformer-based methods.

In single-stage methods, region proposals and object categories are simultaneously obtained through joint encoding. YOLO [21] splits input images into multiple grids. For each grid, it predicts several bounding boxes relative to the grid center. However, it has low detection performance for small or blocked and the object number in each cell is limited. In subsequent versions [22], [23], [24], [25], a series of optimizations have been made for these problems. SSD-series [26], [27], [28], [29] adopts anchor sets and performs detection on feature maps of different resolutions.

In two-stage methods [30], [31], [32], [33], proposal selection is performed first, and then classification and regression of region proposals are performed. Faster-RCNN [33] designs a Region Proposal Network (RPN) to automatically generate proposals inside the network. RPN is followed by the feature-extracted backbone to predict object bounding boxes and scores. Multi-stage methods [34] are to repeat the steps of two-stage methods multiple times and iteratively revise proposals.

The introduction of anchors successfully improves the performance of object detectors. However, it requires a large number of carefully set anchors to cover as many ground-truth labels as possible. Therefore, anchor-free methods are proposed to address this issue. CornerNet [35] proposes to locate objects by detecting the upper left and lower right corners of bounding boxes. On this basis, CenterNet [36] uses the center point to represent targets and obtains bounding boxes by predicting the offset, width, and height of the target center point. With the powerful feature extraction ability of Transformer, ViT [37] first introduces Transformer into computer vision for classification. Based on ViT, DeTR [38] proposes a transformer encoder-decoder architecture for detection.

C. Image Dehazing

Image dehazing is a computer vision technology to remove bad visual effects caused by haze and haze under bad weather conditions. Traditional methods are limited by manual features. Recently, researchers tend to use deep learning-based methods to restore images quickly and reliably. There are divided into two types: supervised and unsupervised.

Supervised methods mostly require prior knowledge, *e.g.*, haze-free labels, depth map, transmission map, atmospheric light, *etc*. DehazeNet [39], Two-layer Gaussian Regression [40], Kernel Regression [41], MSCNN [42], AOD-Net [13], GFN [43] and so on, have good performance and calculation speed. However, training a specific model requires a large number of paired data, which makes it difficult to remove haze in real scenes. Without haze-free labels, unsupervised methods are introduced. They are mostly based on domain translation [44], [45] or image decomposition [46]. Compared with supervised methods, the performance of unsupervised methods is reduced.

D. Attention Mechanisms

Human attention usually concentrates on regions of interest in the whole scene. Inspired by this, attention mechanisms are introduced into computer vision to guide neural network learning. Attention is treated as a dynamic weight adjustment calculation [47] and attention weights are generated by features in the network. Thus, attention mechanisms are sometimes called self-attention. Attention mechanism has been applied in many fields, such as classification, detection, segmentation, video analysis, 3-D vision, and so on. Till now, attention approaches for 2D images are divided into four categories: channel attention, spatial attention, channel with spatial attention, and branch attention.

Channel attention [48], [49], [50] generates masks for each channel of feature maps. It aims at selecting positive channels for prediction. Similar to it, spatial attention [37], [51], [52], [53] generates masks for each position in the entire feature maps. It is used to highlight vital spatial regions. Channel with spatial attention [54], [55], [56] combines above two attention and gets better performance. Branch attention [57], [58], [59], [60] generates masks for different branches and fuse all features to one feature.

The attention module in our paper aims at combining hazy and dehazing features from two branches by self-attention. Thus, we design a branch attention method, which embeds height and width positional information into channel attention with few additional parameters.

III. Methodology

In this section, we first describe details of our proposed BAD-Net for the joint task of dehazing and detection. Next, we introduce a training strategy for further improve BAD-Net performance, as shown in Fig. 2.

A. Dehazing Module

The atmospheric scattering model [61] is often used to generate hazy images, as shown in (1)

$$I(x) = J(x)t(x) + A(1 - t(x)). \tag{1}$$

Here, $t(x)$ is the corresponding transmission map, A is the global atmospheric light, $I(x)$ is the hazy image and $J(x)$ is the corresponding hazy-clean image. Thus, $J(x)$ is obtained by (2)

$$J(x) = \frac{1}{t(x)} I(x) - A\frac{1}{t(x)} + A. \tag{2}$$

Here, we adopt the same approach as AOD-Net [13], which unifies the two changeable parameters $t(x)$ and A into one variable ω, which is defined as (3)

$$\omega = \frac{\frac{1}{t(x)}(I(x) - A) + (A - b)}{I(x) - 1}. \tag{3}$$

Then, ω is substituted into (2), b is the constant bias with value 1. $J(x)$ is re-defined as (4)

$$J(x) = \omega I(x) - \omega + b. \tag{4}$$

From this point of view, ω fully depends on the hazy image $I(x)$, so that it can be modeled by a convolutional neural network. The

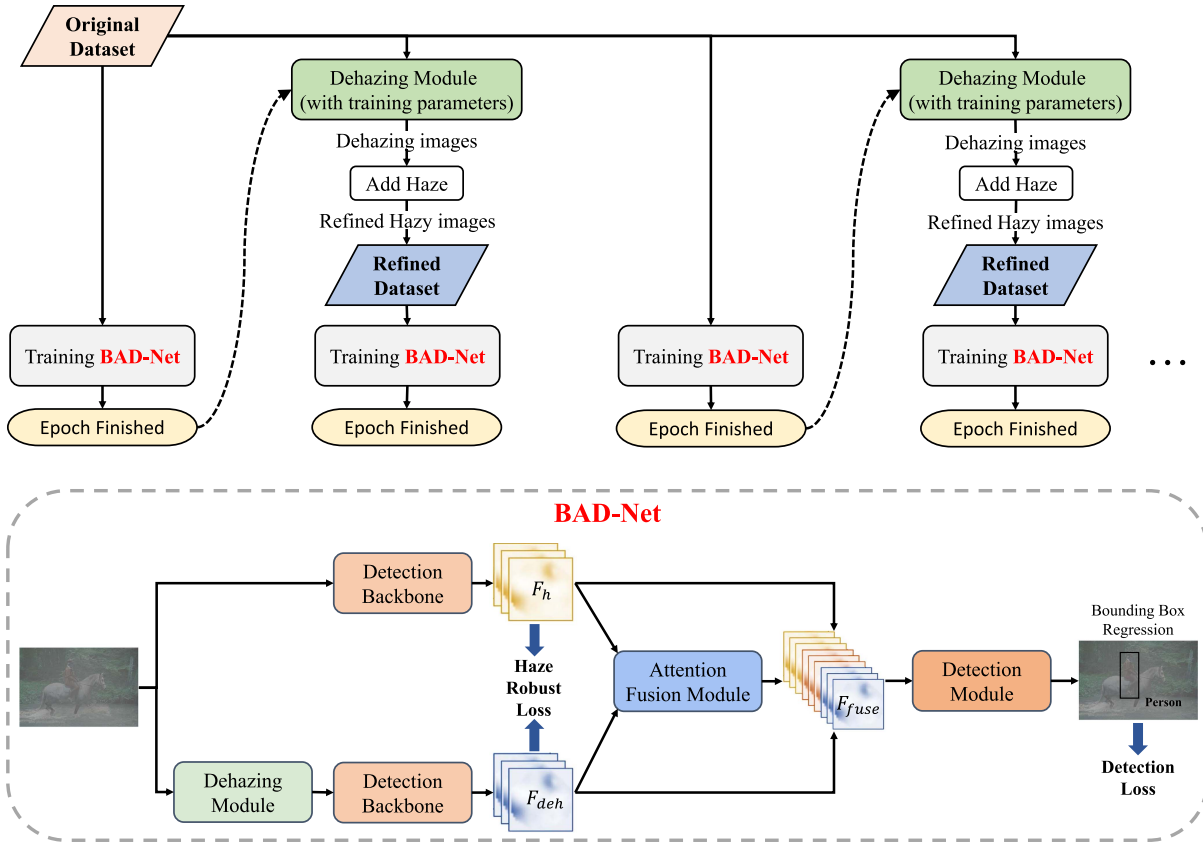

Fig. 2. The architecture of proposed BAD-Net and an interval iterative training strategy for BAD-Net. BAD-Net adopts a bilinear branch structure. One branch directly extracts the hazy features, while the other branch extracts the dehazing features processed by the dehazing module. An attention fusion module is designed to extract the fusion features. After that, the spliced features are sent to the detection module for bounding box prediction. To further improve detection-friendly dehazing, BAD-Net is iteratively trained on the original hazy and refined dataset at one-epoch intervals. The refined dataset is generated by the dehazing module with stable parameters at the previous epoch.

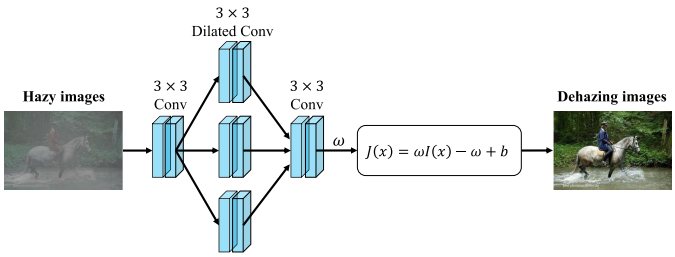

Fig. 3. The architecture of our dehazing module. Based on AOD-Net, this module is implemented by convolutions of different dilation rates.

hazy degradation varies by signals, modeled by a static mapping function difficultly. This dehazing module is input-adaptive for multi-level hazy images.

Different from the original implementation, we extract multi-scale features through three-column dilated convolution. On the premise of not losing resolution, dilated convolution expands the receptive field of convolution kernels and locates targets more accurately. Architecture of the dehazing module is shown in Fig. 3. First, the hazy image is input into two 3×3 convolutions with dilation rate 1 (traditional convolution). Then, feature maps enter three columns of dilated convolution with different dilation rates. The first column is two convolutions with dilation rate 1. The second column is two convolutions with dilation rates 1 and 2. The third column is two convolutions with dilation rate 2. After being aggregated by channel dimension, spliced features are entered into a 3×3 and 1×1 convolution with dilation rate 1. Finally, ω is generated by hazy image $I(x)$. Through (4), dehazing images are obtained.

B. Detection Module

We choose a two-stage detector Faster-RCNN [33] as the detection module. MobilenetV3-large [62] is utilized as a light-weight backbone to extract feature maps, which is pre-trained on ImageNet [63]. Then, feature maps are input into Region Proposal Network to automatically generate proposals for the input image. Faster-RCNN is easy to be extended and optimized. As shown in Fig. 4, we adopt the same architecture and loss functions as the original Faster-RCNN.

C. Attention Fusion Module

We design a hazy-aware attention fusion module for combining hazy and dehazing features from two branches. Dehazing features may lead to bad impacts when the dehazing module performs poorly. Thus, this fusion module is designed to solve

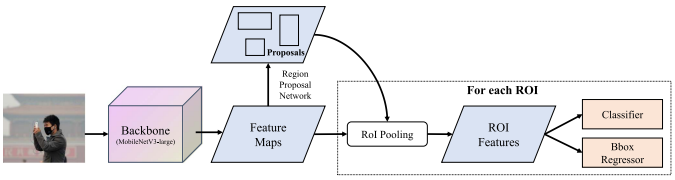

Fig. 4. The architecture of our detection module. We employ the Faster-RCNN detection model with a pre-trained MobileNetV3-large backbone.

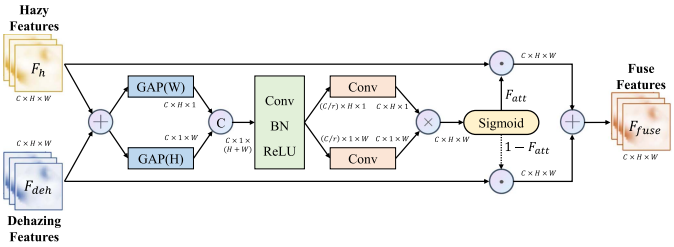

Fig. 5. The architecture of our designed attention fusion module. Attention maps are generated by operating separately in the H and W dimensions. We obtain the final fusion feature, which is added by multiplying the attention map with the original hazy and dehazing feature.

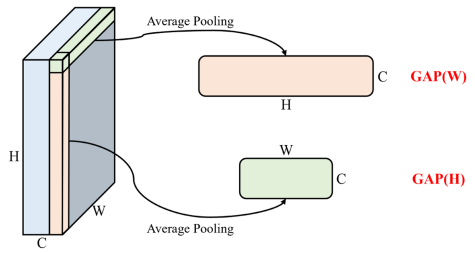

Fig. 6. Flowchart of pooling in attention fusion module. We perform average pooling on height and width dimensions respectively to retain spatial context.

the semantic inconsistency between the dehazing features and the original hazy features. Most attention methods only depend on feature channels. Attention features are often $C \times 1 \times 1$, with one value representing each channel. They may lose spatial information of feature maps, which is critical for generating discriminative and selective attention maps of target locations. Thus, average pooling is performed on height and width dimensions respectively to retain spatial context, as shown in Fig. 5.

Specifically, we operate point-wise addition on haze and dehazing features, to obtain a fused feature. This feature is average pooled on W and H dimensions, separately. The operation is shown in Fig. 6. These two pooled features are spliced to get a $C \times 1 \times (H + W)$ feature. The splicing feature enters a convolution with kernel 1×1, batch normalization, and a ReLU activation function, sequentially. These operations are denoted as (5)

$$F_{tmp} = f_{relu}(f_{BN}(f_{Conv}([Pool^W(X), Pool^H(X)]))). \quad (5)$$

Here, F_{tmp} is $(C/r) \times 1 \times (W + H)$ by convolution f_{Conv}.

Then, F_{tmp} is divided into two features of $(C/r) \times H \times 1$ and $(C/r) \times 1 \times W$. These two features are entered into a convolution $f_{Conv'}$ with kernel 1×1, respectively. The channel of features is changed from (C/r) to C again. We perform the obtained features $C \times H \times 1$ and $C \times 1 \times W$ by matrix multiplication. An attention map F_{att} $(C \times H \times W)$ is generated through a Sigmoid activation function. These operations are denoted as (6)

$$F_{att} = \delta(f_{Conv1'}(F^H_{tmp}) \otimes f_{Conv2'}(F^W_{tmp})). \quad (6)$$

Finally, we obtain the fuse feature F_{fuse} by combining the attention map F_{att} with both haze feature F_h and dehazing feature F_{deh}. These operations are denoted as (7)

$$F_{fuse} = (F_h \odot F_{att}) \oplus (F_{deh} \odot (1 - F_{att})). \quad (7)$$

Channel attention only focuses on information fusion between different channels. Based on this, we also considers both spatial and channel context. We encode the target location information through separate operations of height and width dimension. The generated attention map better represents positions of targets for further detection.

D. Loss Function

IA-YOLO [5] proves that adding image restoration loss to joint training achieves slower convergence and worse detection performance. The reason is that restoration loss focuses on the quality of restored images. Meanwhile, detection loss focuses on the extraction of target features in regions of interest rather than background. Detection aims at extracting the features that are insensitive to high abstraction. Yet, dehazing aims at extracting features that are sensitive to detail and low abstraction. Therefore, there is a conflict between two losses. The entire model may converge to a local optimum point and cannot reach the global one.

From this point of view, in addition to detection loss, we also introduce a self-supervised loss for guiding the learning direction of dehazing module, called Haze Robust Loss (HR Loss). It is to achieve detection-friendly dehazing. As shown in Fig. 2, we obtain the feature maps F_h and F_{deh} from detection backbone in first and second branch, respectively. Global average pooling are then performed on these feature maps, to obtain representative vectors of $C \times 1 \times 1$ size. Finally, HR Loss is computed by point-wise KL-divergence, denoted as (8).

$$L_{hr} = GAP(F_h) \cdot \log \frac{GAP(F_h)}{GAP(F_{deh})}. \quad (8)$$

In knowledge distillation [64], student network learns the probability distribution of teacher network, and KL divergence is used to measure the difference between two distributions. Inspired by this, we use it as a soft constraint to make dehazing features similar to original features. Above all, we treat two branches as two distortions of input images in self-supervised learning, as shown in Fig. 7. We aim to learn a detection model, which is robust to different haze levels. This can be achieved by constraining the similarity of haze features and dehazing features extracted by the detection backbone. Restoration loss learns image quality restoration including the background information. This may introduce noise and harm further detection. Rather than this, we move the dehazing constraint forward to the detection

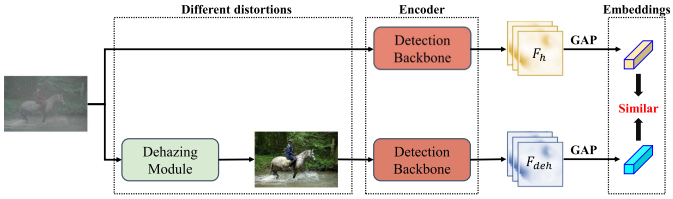

Fig. 7. The flowchart of Haze Robust Loss. Two branches are regarded as two distortions of input images according to self-supervised learning. HR Loss constraints the similarity between features with self supervision, so as to make the detection module more robust.

extraction part. It helps to make the dehazing module pay more attention to the learning of target region features.

Above all, total loss of BAD-Net is denoted as (9)

$$L_{total} = L_{det} + \alpha L_{hr}. \tag{9}$$

Here, L_{det} is the detection loss of proposals and detections in original Faster-RCNN. α is used to balance the importance of L_{det} and L_{hr}. It is experimentally set via ablative analysis in Section IV-D3.

E. Interval Iterative Training Strategy

Above all, we propose an end-to-end union detection network (BAD-Net) with weakly-supervised dehazing. The extraction backbone of the detection module shares the same parameters in two branches. To further improve the accuracy of bounding boxes for detection, we introduce an interval iterative training strategy, as shown in Fig. 2. Inspired by a denoising method IDR [65], we update restoration loss to detection loss for weakly-supervised dehazing and optimize this method using interval approaches to achieve detection training stability.

To further strengthen dehazing ability without restoration loss, we conduct self-supervised constraints on the dehazing module at the training data level. BAD-Net is iteratively trained on the original hazy dataset and refined datasets at one-epoch intervals. Different from the original dataset, refined datasets are generated by the dehazing module with training parameters at the end of the previous epoch.

At even epochs, a BAD-Net model is trained on the original hazy dataset X. At the epoch end, we obtain a dehazing module f_{dh} with stable training parameters for generating refined datasets in the next epoch. At odd epochs, original hazy images are first input into f_{dh} to generate dehazing images $f_{dh}(X)$. Then, we apply an algorithm g_h (Algorithm 2) for adding haze of different levels to these dehazing images. Through this, a refined hazy training dataset $g_h(f_{dh}(X))$ is obtained. By this means, the detection module directly learns the image distribution and bias generated by the dehazing module. Thus, it can avoid the adverse effects of random noise by dehazing. Compared with the original hazy dataset, the entire model trained on refined datasets has a stronger generalization ability on actual hazy images.

Most importantly, we adopt the original hazy dataset and refined datasets alternately during training. The dehazing module in BAD-Net has no direct restoration loss constraint. When the dehazing module has poor performance, the generated refined datasets may have a large distribution deviation from the original dataset. This will lead to difficulty in the subsequent training process. Thus, we use the original hazy dataset to fine-tune model parameters to achieve the purpose of buffering.

Algorithm 1: Interval Iterative Training Strategy for Refined Datasets.

Input: Original hazy images X, haze algorithm g_h, Total epoch N.
Output: Final BAD-Net model P^N.

Initialize BAD-Net model P^0, in addition to feature extraction backbone of the detection module with pretrained parameters on ImageNet;
foreach *n in range(N)* **do**
  if n % 2 == 0 **then**
    Training dataset is original hazy dataset;
    Optimize current model P^n by minimizing total loss for one epoch;
  else
    Initialize a new dehazing model f_{dh}^{n-1};
    Create new refined training dataset $g_h(f_{dh}^{n-1}(X))$;
    Training dataset is refined hazy dataset;
    Optimize current model P^n by minimizing total loss for one epoch;

Algorithm 2: Adding Haze on a Dehazing Image.

Input: dehazing image DI, Brightness A, haze thickness $level$.
Output: Generated hazy image HI.

(H, W, _) = DI.shape;
HI = DI.copy();
foreach *r in range(H)* **do**
  foreach *c in range(W)* **do**
    $d = -0.04 * \sqrt{(r - H//2)^2 + (c - W//2)^2}) + \sqrt{max(H, W)}$;
    $td = math.exp(-level * d)$;
    $HI[r][c] = HI[r][c] * td + A * (1 - td)$;

To achieve fast refined dataset iteration, we design a training strategy as Algorithm 1. This strategy makes the dehazing module converge faster and achieve better performance. In the last stage of training, the models trained on the original and refined dataset achieve approximately equal verification results.

The algorithm g_h of adding haze is shown as Algorithm 2. (1) is also expressed as (10). The algorithm g_h is base on (10)

$$I(x) = J(x)e^{-\beta d(x)} + A(1 - e^{-\beta d(x)}), \tag{10}$$

where β is the atmospheric scattering coefficient. $d(x)$ is the corresponding depth map, which is simulated by calculating the euclidean distance between each pixel value and the center point. Besides, we set $A = 0.5$ and $level = 0.05 + 0.01 * i, i = 0, 1, 2, \ldots, 9$. It means ten different levels of haze are randomly added to each dehazing image.

TABLE I
DETAILED STATISTICAL INFORMATION ON RTTS AND VOCHAZE

	RTTS	VOCn-tv (VOCh-tv)	VOCn-test (VOCh-test)
Person	11,366	15,576	5,227
Bicycle	698	1,208	389
Motorbike	25,317	1,141	369
Bus	2,590	909	254
Car	1,232	4,008	1,541
Image	4,322	8,383	2,795
Total	41,203	22,842	7,780

IV. EXPERIMENTS

In this section, we first introduce two experimental datasets, VOChaze and RTTS. Then, we introduce the implementation details of BAD-Net on these datasets. In particular, we evaluate our method on a synthetic dataset and a real-world dataset to compare with SOTA methods. Moreover, ablation studies are also conducted to further validate the validity of our network.

A. Datasets

Few public detection datasets consist of hazy conditions. Therefore, we choose a real-world hazy dataset, RTTS. Besides, we choose a synthetic hazy dataset based on VOChaze. The statistical details are shown in Table I.

RTTS: RTTS [66] is the largest annotated detection dataset in hazy conditions. It contains 4322 real-world hazy images, mostly in traffic scenes. There are five categories, such as person, bicycle, motorbike, bus, and car. 41203 bounding boxes are labeled. Among them, 11606 boxes are marked as "difficult".

VOChaze: This synthetic dataset is introduced in [5]. An algorithm for adding haze is performed on the VOC dataset, as shown in Algorithm 2. Before each training image is input into the model, a $level$ value is randomly selected, and then haze is added to the clear image. Algorithm 2 is executed by $2/3$ probability each time. There are four main sub-datasets. VOCn-tv and VOCn-test are the trainval split of VOC07+VOC12 and the test split of VOC07, respectively. VOCh-tv and VOCh-test are synthetic hazy datasets, based on the former two. It is noted that only five types of data (same as RTTS) are filtered into the final datasets.

B. Implementation Details

During training, only random horizontal flip is applied to enhance input images. The whole model is trained using mini-batch Adam [67] optimizer with initial rate of 10^{-4}, and weight decay of 10^{-4}. The batch size is set to 8 and the max epoch is 50. In the 30th epoch, the learning rate decreased to 10%. α in (9) is set to 0.001 due to multiple experiments. Our experiments are based on the torchvision detection framework of Pytorch. We train and test models on one NVIDIA 3090 GPU.

C. Comparison With the SOTAs

We choose Faster-RCNN with MobilenetV3-large backbone as the baseline. As introduced in Section II-A, we divide the existing methods into the following three types according to strategies: (1) Direct: hazy images are directly trained on detection models; (2) Separate: hazy images are first trained on dehazing models, then input into detection baseline pre-trained on VOCn-tv dataset. Note that there is no training on detection models; (3) Union: hazy images are simultaneously trained on the union model of dehazing and detection model. When the training data includes both VOCn-tv and VOCh-tv, it means that this method needs a haze-free restoration ground truth corresponding to the hazy image. In this paper, for quantitative comparison, we change the detection model in other methods to the baseline.

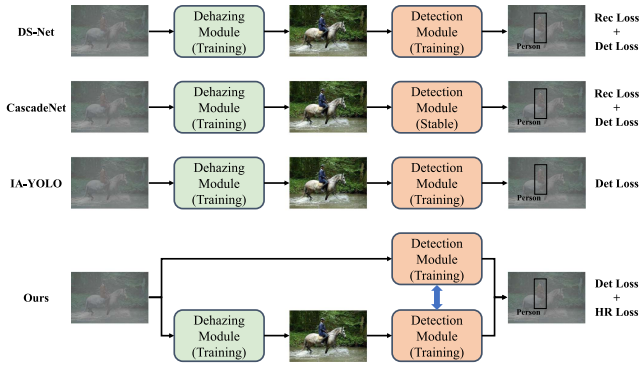

Fig. 8. The introduction of methods with union strategy. Row 1: Models jointly train dehazing module and detection module with restoration and detection loss. Row 2: Detection module is stable with parameters pretrained on common datasets. Models only train dehazing module with two losses. Row 3: Models jointly train two modules with only detection loss. Row 4: Two-branch BAD-Net trains the entire model with only detection loss.

CascadeNet was originally used for denoising. In the implementation, its denoising part is changed to AOD-Net dehazing network. It only updates the parameters of the enhancement module in training using both restoration loss and detection loss. Its detection module is with stable parameters pre-trained on normal no-degraded datasets. The architecture of CascadeNet is shown in the second row of Fig. 8. DS-Net is similar to the first row of Fig. 8. The difference is that the input of its dehazing network is feature maps extracted by the detection extracted backbone. DS-Net used AOD-Net as its dehazing network. IA-YOLO is shown in the third row of Fig. 8. It only uses detection loss to train the dehazing and detection modules from scratch.

As shown in Table II, we compare the mean average precision of the total five classes (mAP) results between state-of-the-art methods. BAD-Net converges stably in a few epochs and achieves the best detection performance in normal and hazy conditions. There are four main observations:

- The performance of methods using the separate strategy greatly depends on its dehazing model. Compared with Baseline-1, the detection performance of AOD-Net and MSBDN is slightly degraded while GridDehaze is improved. This means that the dehazing network may introduce random noise harmful to the detection network. It is because their loss constraints and evaluation metrics are different, and there is no obvious positive correlation between them.

TABLE II
PERFORMANCE COMPARISON WITH STATE-OF-THE-ART METHODS ON HAZY DATASETS

Strategy	Method	Module		Train Dataset	Test Dataset		
		Dehazing	Detection		VOCn-test	VOCh-test	RTTS
Direct	Baseline-1	/	Faster-RCNN	VOCn-tv	82.03	77.77	46.69
	Baseline-2	/	Faster-RCNN	VOCh-tv	81.60	80.96	47.03
Separate	Sep-1	AODNet [13]	Faster-RCNN	VOCh-tv, VOCn-tv	/	73.67	40.59
	Sep-2	MSBDN [68]	Faster-RCNN	VOCh-tv, VOCn-tv	/	76.35	43.90
	Sep-3	GridDehaze [69]	Faster-RCNN	VOCh-tv, VOCn-tv	/	78.11	45.99
Union	CascadeNet [3]	AODNet	Faster-RCNN	VOCh-tv, VOCn-tv	/	75.97	43.09
	DS-Net [4]	Own	Faster-RCNN	VOCh-tv, VOCn-tv	83.07	80.95	50.68
	IA-YOLO [5]	Own	YOLOv3	VOCh-tv	83.46	82.10	50.03
	IA-YOLO [5]	AODNet	Faster-RCNN	VOCh-tv	80.46	78.10	46.33
	BAD-Net (Ours)	AODNet	Faster-RCNN	VOCh-tv	85.86	85.58	53.15

The best result in each column is in red, and the second is in blue.

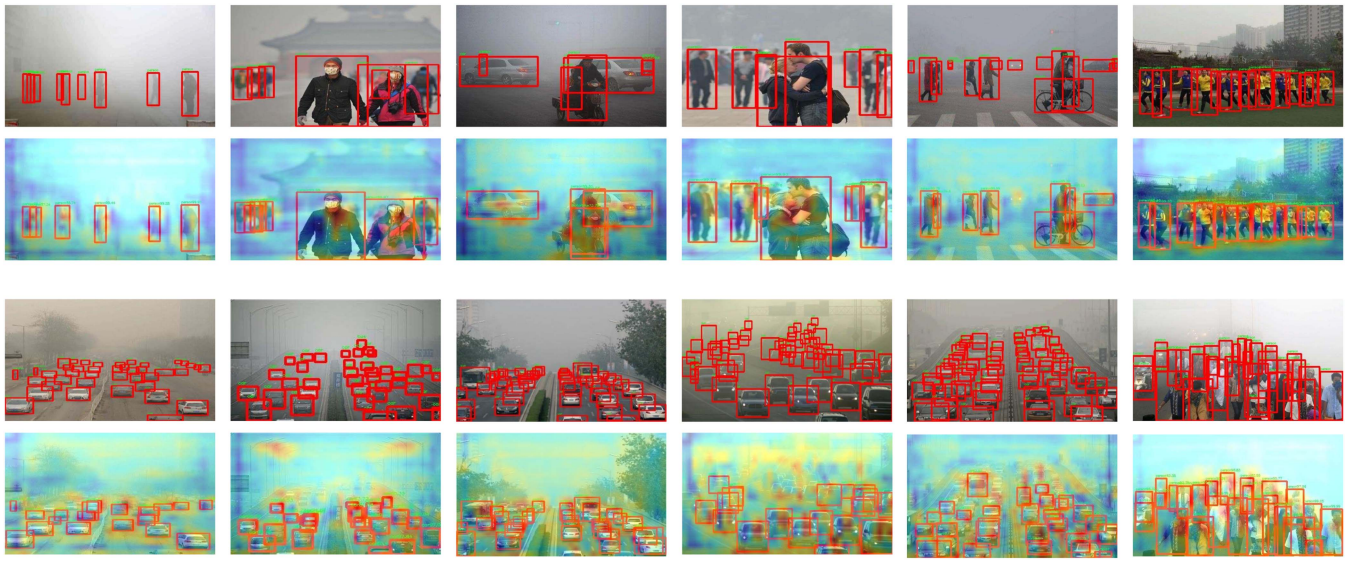

Fig. 9. Some "normal" and "difficult" visualization samples of BAD-Net detection results on a real-world hazy dataset RTTS. Row 1 and 3 are original images with detection labels. Row 2 and 4 are the corresponding class activation maps of the detection module in BAD-Net with predicted results.

- CascadeNet converges slowly and has poor performance. The whole model with two losses is difficult to jump out of the local optimum point during the training process.
- DS-Net performs better, due to its detection model pre-trained on the common MS-COCO dataset. During the whole training process, the performance of DS-Net is gradually declining and does not play a positive correlation joint optimization effect.
- IA-YOLO achieves high accuracy by using only one detection loss. However, when we change its dehazing module to AOD-Net, its training loss is difficult to converge and its performance decreases greatly. It is because the dark channel dehazing algorithm and many traditional digital image processing methods in IA-YOLO are beneficial to detection. This is not a robust framework for detection performance improvement in hazy conditions.

Considering that the dehazing module brings harmful noise to the detection module, BAD-Net uses the double branch method to reduce the impact of harmful features. It is a robust framework for connecting dehazing and detection, *i.e.*, low-level image processing, and high-level pattern recognition. Some samples of BAD-Net detection results are shown in Fig. 9. Our model accurately recognizes targets in the vicinity, even if they are obscured by haze. For "difficult" samples, our model performs poorly on distant small, and heavily occluded objects. In the following ablation study, we will analyze the impact of each module.

D. Ablation Study

In this subsection, we explore our method BAD-Net with several different settings on two experimental datasets. There are four ablation studies, such as different module combinations (Section IV-D1), different attention mechanism (Section IV-D2), different loss weights (Section IV-D3), different dehazing and detection modules (Section IV-D4), and efficiency analysis for inference speed (Section IV-D5).

TABLE III
PERFORMANCE COMPARISON WITH DIFFERENT MODULE COMBINATIONS ON VOCHAZE AND RTTS

Two Branch	L1 Loss	L2 Loss	HR Loss	Training Strategy	VOCn-test	VOCh-test	RTTS
					80.46	78.10	46.33
✓					83.07	82.15	49.43
✓	✓				82.91	82.06	48.98
✓		✓			83.76	83.85	50.23
✓			✓		84.51	84.36	51.32
✓			✓	✓	85.86	85.58	53.15

The best result in each column is in red, and the second is in blue.

1) Module Ablation: An ablation study is executed to analyze the performance of each module in BAD-Net. Here, we divide the entire model into three sub-modules: two-branch structure, HR Loss, and training strategy. The experimental results are shown in Table III. Compared with the single branch in traditional methods, the two-branch model performs better and gets +3.1 mAP performance on RTTS dataset.

For HR Loss, there are three variants: L1 loss, L2 loss, and KL-divergence loss. The performance of L1-based model decreases by 0.16, 0.09, and 0.45 mAP on three datasets. This is because L1 loss is a strong constraint. It makes the detection loss more difficult to converge, resulting in lower accuracy. Compared with L1 loss, L2 loss is a relatively weak constraint. It does not affect the convergence of detection loss, so there is a slight performance improvement of 0.69, 1.7, and 0.8 mAP on three datasets. KL-divergence loss is used to measure the difference between the two characteristic distributions. Minimizing the KL divergence is equivalent to maximizing the likelihood ratio [70]. It focuses on the probability distribution of features and weakens the influence of background information to a certain extent. The haze robust loss based on KL divergence effectively improves the anti-haze capability of the detection backbone which generates detection-friendly features. Therefore, it pays more attention to the target area information, which is consistent with the ideas proposed in this module. It obtains +1.44, +2.21, and +1.89 improvement and performs better than L1-based and L2-based models.

With the proposed training strategy, the model further achieves +1.35, +1.22, and +1.83 mAP improvement. Above all, these sub-modules have brought performance improvement to the entire model, which proves their effectiveness.

2) Attention: We explore different attention mechanisms in Attention Fusion Module. Here, we choose two branch attention methods as the baselines: SK-Net [57] and AFF [60]. Based on SK-Net, AFF adopts two branches to extract global and local attention features. Using convolutions rather than fully connected layers reduces the number of parameters. Our method performs average pooling on height and width dimensions respectively to enhance the target location information. Due to this characteristic, further object detection shows better performance than other attention-based models. The detection result comparison is shown in Fig. 10 and Table IV. Compared with AFF, our attention fusion module obtains +0.45, +0.51, and +0.26 mAP improvement.

3) Loss Weight: We explore the optimal balance between detection loss and HR Loss, *i.e.*, the α value in (9). A comparison of results is shown in Fig. 11. The smaller α is, the faster BAD-Net converges. The model performs best when α is set to 0.001. It indicates that HR Loss slows down the over-fitting during training process to some extent. HR Loss also increases the generalization of the entire model. However, when α is rather small, HR Loss is unnecessary relative to the detection loss. It does not serve as a valid constraint. During training, the total loss fluctuates greatly and eventually converges to a large loss value. As a result, poor prediction performance is produced.

4) Dehazing and Detection Module: We compare BAD-Net with different detection modules. The results are shown in Table V. We choose following common detection models: Faster-RCNN (MobileNetV3-large), Faster-RCNN (Resnet50), FCOS, RetianNet and SSD. The detection module is vitally important to the final performance in this task. According to parameters of the detection module, the performance is approximately related to it. Faster-RCNN (Resnet50) is similar to FCOS and

(a) Original images with labels (b) Detection results with SK-Net (c) Detection results with AFF (d) Detection results with our method

Fig. 10. Some detection visualization results of BAD-Net with different attention mechanism on RTTS. Compared with the other two methods, our method brings a larger performance improvement for subsequent detection.

TABLE IV
PERFORMANCE COMPARISON WITH DIFFERENT ATTENTION MECHANISMS ON VOCHAZE AND RTTS

Model	VOCn-test	VOCh-test	RTTS
SK-Net	84.90	84.66	52.45
AFF	85.41	85.07	52.89
Ours	**85.86**	**85.58**	**53.15**

Bold font highlights the best results in each column.

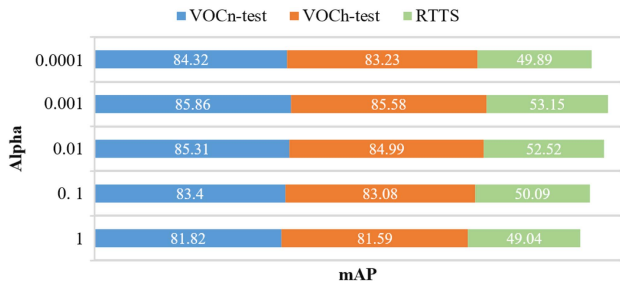

Fig. 11. Performance comparison with different loss weights between Haze Robust Loss and detection loss on VOCHaze and RTTS.

TABLE V
PERFORMANCE COMPARISON WITH DIFFERENT DETECTION MODULES ON VOCHAZE AND RTTS

Model	VOCn-test	VOCh-test	RTTS
Faster-RCNN (Baseline)	85.86	85.58	53.15
Faster-RCNN (Resnet50)	85.91	85.60	53.54
FCOS (Resnet50)	**85.90**	**85.61**	**53.56**
RetinaNet (Resnet50)	85.90	85.58	53.53
SSD300 (VGG16)	83.08	82.71	52.29

Bold font highlights the best results in each column.

TABLE VI
PERFORMANCE COMPARISON WITH DIFFERENT DEHAZING MODULES ON VOCHAZE AND RTTS

Model	VOCn-test	VOCh-test	RTTS
AOD-Net	85.86	85.58	53.15
MSBDN	**86.01**	**85.89**	**53.21**
GridDehaze	85.88	85.63	53.19

Bold font highlights the best results in each column.

TABLE VII
EFFICIENCY COMPARISON WITH DIFFERENT MODELS ON RTTS

Model	Params	Speed (ms)	mAP
Faster-RCNN	19M	**41**	47.03
AOD-Net+Faster-RCNN	+17K	45	40.59
MSBDN+Faster-RCNN	+31M	114	43.90
GridDehaze+Faster-RCNN	+958K	86	45.09
IA-YOLO (Faster-RCNN)	+165K	67	50.03
BAD-Net	+340K	72	**53.15**

Bold font highlights the best results in each column.

RetinaNet, all better than Faster-RCNN (MobileNetV3-large). However, they have much more training and inference time. SSD (VGG16) is the single-stage model, and its performance is worse than that of two-stage models.

Besides, we compare BAD-Net with different dehazing modules. The results are shown in Table VI. Compared with AOD-Net, MSBDN and GridDehaze have slightly improved. This is because they have more complex structures and more trainable parameters.

5) Efficiency Analysis: The parameter amount and inference speed are shown in Table VII. These methods all use a Faster-RCNN detection network with a MobilenetV3-large backbone. When testing the inference speed, we use one NVIDIA 3090 GPU to test on the test images of the RTTS dataset, that is, images with $848 \times 480 \times 3$ resolution. Compared with IA-YOLO, BAD-Net adds 175 K trainable parameters and 5 ms inference time. While real-time performance is ensured, the detection accuracy of BAD-Net has also been improved.

V. CONCLUSION

In this paper, we present a novel bilinear attention detection network BAD-Net for hazy conditions. Through the designed attention fusion module, we fully integrate both haze and dehazing features. This effectively improves the complementarity and richness of target features. Besides, we also introduce haze robust loss and interval iterative data refinement strategy to weakly-supervised restrict the dehazing module learning. BAD-Net doesn't need image restoration labels so that it can be applied to real scenarios. The experimental results show that BAD-Net achieves the highest detection accuracy on a real-world hazy dataset RTTS and a synthetic hazy dataset VOChaze. BAD-Net converges stably and fastly during the training process. It has few parameters and a high inference speed for real-time detection. Most importantly, it is a powerful framework for connecting low-level image enhancement and high-level vision tasks, in which each module is easy to be replaced and extended.

In the future, we intend to explore a union solution that positively correlates detection loss with restoration loss by using reinforcement learning. Besides, it is also a valuable research direction to design a restoration loss that only judges the quality of target regions.

REFERENCES

[1] C. Chen, Z. Xiong, X. Tian, Z.-J. Zha, and F. Wu, "Real-world image denoising with deep boosting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 12, pp. 3071–3087, Dec. 2019.
[2] Z. Cui et al., "Exploring resolution and degradation clues as self-supervised signal for low quality object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 473–491.
[3] D. Liu, B. Wen, J. Jiao, X. Liu, Z. Wang, and T. S. Huang, "Connecting image denoising and high-level vision tasks via deep learning," *IEEE Trans. Image Process.*, vol. 29, pp. 3695–3706, 2020.
[4] S.-C. Huang, T.-H. Le, and D.-W. Jaw, "DSNet: Joint semantic learning for object detection in inclement weather conditions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 8, pp. 2623–2633, Aug. 2021.
[5] W. Liu, G. Ren, R. Yu, S. Guo, J. Zhu, and L. Zhang, "Image-adaptive yolo for object detection in adverse weather conditions," 2021, *arXiv:2112.08088*.
[6] N. Akhtar, M. A. Jalwana, M. Bennamoun, and A. Mian, "Attack to fool and explain deep networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 5980–5995, Oct. 2021.
[7] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Image quality assessment: Unifying structure and texture similarity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2567–2581, May 2022.
[8] J. Yim and K.-A. Sohn, "Enhancing the performance of convolutional neural networks on quality degraded datasets," in *Proc. Int. Conf. Digit. Image Comput.: Techn. Appl.*, 2017, pp. 1–8.
[9] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," 2019, *arXiv:1903.12261*.
[10] M. Singh, S. Nagpal, R. Singh, and M. Vatsa, "Dual directed capsule network for very low resolution image recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 340–349.
[11] C. Wang, C. Li, B. Luo, W. Wang, and J. Liu, "RiWNet: A moving object instance segmentation network being robust in adverse weather conditions," 2021, *arXiv:2109.01820*.
[12] D. Dai, Y. Wang, Y. Chen, and L. Van Gool, "Is image super-resolution helpful for other vision tasks?," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2016, pp. 1–9.
[13] B. Li, X. Peng, Z. Wang, J. Xu, and D. Feng, "AOD-Net: All-in-one dehazing network," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 4770–4778.
[14] M. W. Gondal, B. Schölkopf, and M. Hirsch, "The unreasonable effectiveness of texture transfer for single image super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 80–97.
[15] Y. Pei, Y. Huang, Q. Zou, Y. Lu, and S. Wang, "Does haze removal help CNN-based image classification?," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 682–697.
[16] S. Li et al., "Single image deraining: A comprehensive benchmark analysis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3838–3847.
[17] S. Diamond, V. Sitzmann, F. Julca-Aguilar, S. Boyd, G. Wetzstein, and F. Heide, "Dirty pixels: Towards end-to-end image processing and perception," *ACM Trans. Graph.*, vol. 40, no. 3, pp. 1–15, 2021.
[18] D. Liu, B. Wen, X. Liu, Z. Wang, and T. S. Huang, "When image denoising meets high-level vision tasks: A deep learning approach," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, 2018, pp. 842–848.
[19] T. Son, J. Kang, N. Kim, S. Cho, and S. Kwak, "Urie: Universal image enhancement for visual recognition in the wild," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 749–765.

[20] M. Haris, G. Shakhnarovich, and N. Ukita, "Task-driven super resolution: Object detection in low-resolution images," in *Proc. Int. Conf. Neural Inf. Process.*, 2021, pp. 387–395.
[21] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, 2016, pp. 779–788.
[22] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7263–7271.
[23] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.
[24] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.
[25] G. Jocher, A. Chaurasia, A. Stoken, J. Borovec, and K. NanoCode012, "Ultralytics/yolov5: V6. 1-tensorrt, tensorflow edge tpu and openvino export and inference," 2022, Art. no. 6222936, doi: 10.5281/ZENODO.
[26] W. Liu et al., "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.
[27] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "DSSD: Deconvolutional single shot detector," 2017, *arXiv:1701.06659*.
[28] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.
[29] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10781–10790.
[30] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 580–587.
[31] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.
[32] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.
[33] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.
[34] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6154–6162.
[35] H. Law and J. Deng, "CornerNet: Detecting objects as paired keypoints," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 734–750.
[36] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "CenterNet: Keypoint triplets for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 6569–6578.
[37] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
[38] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.
[39] B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao, "DehazeNet: An end-to-end system for single image haze removal," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5187–5198, Nov. 2016.
[40] X. Fan, Y. Wang, X. Tang, R. Gao, and Z. Luo, "Two-layer Gaussian process regression with example selection for image dehazing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 12, pp. 2505–2517, Dec. 2017.
[41] C.-H. Xie, W.-W. Qiao, Z. Liu, and W.-H. Ying, "Single image dehazing using Kernel regression model and dark channel prior," *Signal, Image Video Process.*, vol. 11, no. 4, pp. 705–712, 2017.
[42] W. Ren, S. Liu, H. Zhang, J. Pan, X. Cao, and M.-H. Yang, "Single image dehazing via multi-scale convolutional neural networks," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 154–169.
[43] W. Ren et al., "Gated fusion network for single image dehazing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3253–3261.
[44] D. Engin, A. Genç, and H. Kemal Ekenel, "Cycle-dehaze: Enhanced cyclegan for single image dehazing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 825–833.
[45] Y. Mo, C. Li, Y. Zheng, and X. Wu, "DCA-CycleGAN: Unsupervised single image dehazing using dark channel attention optimized cyclegan," *J. Vis. Commun. Image Representation*, vol. 82, 2022, Art. no. 103431.
[46] Y. Gandelsman, A. Shocher, and M. Irani, ""Double-DIP": Unsupervised image decomposition via coupled deep-image-priors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 11026–11035.
[47] M.-H. Guo et al., "Attention mechanisms in computer vision: A survey," *Comput. Vis. Media*, vol. 8, pp. 331–368, 2022.
[48] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.
[49] Z. Yang, L. Zhu, Y. Wu, and Y. Yang, "Gated channel transformation for visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11794–11803.
[50] Z. Qin, P. Zhang, F. Wu, and X. Li, "FcaNet: Frequency channel attention networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 783–792.
[51] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.
[52] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 7354–7363.
[53] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 13713–13722.
[54] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.
[55] J. Park, S. Woo, J.-Y. Lee, and I. S. Kweon, "BAM: Bottleneck attention module," 2018, *arXiv:1807.06514*.
[56] J. Fu et al., "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3146–3154.
[57] X. Li, W. Wang, X. Hu, and J. Yang, "Selective Kernel networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 510–519.
[58] H. Zhang et al., "ResNeSt: Split-attention networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 2736–2746.
[59] Y. Chen, X. Dai, M. Liu, D. Chen, L. Yuan, and Z. Liu, "Dynamic convolution: Attention over convolution Kernels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11030–11039.
[60] Y. Dai, F. Gieseke, S. Oehmcke, Y. Wu, and K. Barnard, "Attentional feature fusion," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2021, pp. 3560–3569.
[61] S. K. Nayar and S. G. Narasimhan, "Vision in bad weather," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, 1999, pp. 820–827.
[62] A. Howard et al., "Searching for MobileNetV3," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 1314–1324.
[63] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[64] G. Hinton et al., "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.
[65] Y. Zhang, D. Li, K. L. Law, X. Wang, H. Qin, and H. Li, "IDR: Self-supervised image denoising via iterative data refinement," 2021, *arXiv:2111.14358*.
[66] B. Li et al., "Benchmarking single-image dehazing and beyond," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 492–505, Jan. 2019.
[67] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[68] H. Dong et al., "Multi-scale boosted dehazing network with dense feature fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2157–2167.
[69] X. Liu, Y. Ma, Z. Shi, and J. Chen, "GridDehazeNet: Attention-based multi-scale network for image dehazing," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 7314–7323.
[70] T. Lartigue, S. Bottani, S. Baron, O. Colliot, S. Durrleman, and S. Allassonnière, "Gaussian graphical model exploration and selection in high dimension low sample size setting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 9, pp. 3196–3213, Sep. 2021.

Chengyang Li (Graduate Student Member) received the MS degree in computer technology from the China University of Petroleum, Beijing, China, in July 2020. He is currently working toward the PhD degree in computer software and theory with Peking University. His research interests include image processing, video understanding, and multimodal intelligence. Currently, he mainly focuses on high-level vision model optimization on degraded images. He has published several papers in *Pattern Recognition*, *NeuroComputing*, ICASSP and so on. Besides, he serves as a reviewer for *Pattern Recognition*, *Neural Processing Letters* and *IEEE Transactions on Circuits and Systems for Video Technology*.

Heng Zhou (Graduate Student Member) He is currently working toward the PhD degree in electronic science and technology with Xidian University, Xi'an, China. His current research interests include image processing, pattern recognition, and their applications in infrared target detection and segmentation.

Yang Liu is currently working toward the MS degree in computer science and technology with the Institute of Systems Engineering of AMS, China. His research interests include video communication, image dehazing and object detection in adverse weather conditions.

Caidong Yang received the graduate degree from the Beijing University of technology, Beijing, in 2020. He is currently working towrad the MS degree in information and communication engineering with the Institute of Systems Engineering of AMS, China. His research interests include super-resolution reconstruction and object detection.

Yongqiang Xie has been engaged in the research of image, video and communication technology for a long time, and has made outstanding contributions in signal processing. Now, he is a leader of the dedicated algorithm group for Chinese video and audio standards, a executive deputy director of Rich Media Committee in Chinese Institute of Command and Control, and a member of China Post-Doctoral Fund Review Committee. He is a researcher of Institute of Systems Engineering at AMS and a professor of School of Systems Science and Engineering at Sun Yat-Sen University. His research interests include image processing, computer vision, multimedia analysis, machine learning, and pattern recognition. He is currently undertaking more than 20 major national scientific research projects. He has published two books, around 50 technical articles in refereed journals and proceedings. He holds more than 100 granted patents. He obtained the Qiu Shi Award, selected as a candidate of the one hundred plus one thousand plus ten thousand talents project of the new century and outstanding mid-aged expert. He was selected as an Expert enjoying the Government Special Subsidy.

Zhongbo Li is mainly dedicated to video understanding and intelligent analysis, including pedestrian detection, crowd counting, and intelligent transportation. He is now a senior engineer of Institute of Systems Engineering. AMS. He has won the second prize of National Science and Technology Progress Award and the first prize of Provincial and Ministerial Science and Technology Award. He has published more than 20 high-level papers in related fields and published one academic book.

Liping Zhu is currently an associate professor with the Computer Technology Department of China University of Petroleum (Beijing). Her research mainly focuses on applications of computer vision, machine learning, and data mining on oil and gas. Meanwhile, she is an administrator of Beijing Key Laboratory of Petroleum and Data Mining. She presides or undertakes six projects such as 863, sub-projects of major national oil and gas special projects, Beijing Municipal Science and Technology Commission Ladder Plan, and so on. Besides, she pays attention to the application of academic algorithms in practical scenarios and undertakes 10+ projects entrusted by enterprises. In the field of security monitoring and petroleum, she has published more than 20 high-level journal papers, several patents, and software copyrights.