DOI: 10.1111/cgf.14692
Pacific Graphics 2022
N. Umetani, E. Vouga, and C. Wojtan
(Guest Editors)

Volume 41 (2022), Number 7

TogetherNet: Bridging Image Restoration and Object Detection Together via Dynamic Enhancement Learning

Yongzhen Wang[1], Xuefeng Yan[1†], Kaiwen Zhang[1], Lina Gong[1], Haoran Xie[2], Fu Lee Wang[3], Mingqiang Wei[1,4]

[1]School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China
[2]Department of Computing and Decision Sciences, Lingnan University, Hong Kong, China
[3]School of Science and Technology, Hong Kong Metropolitan University, Hong Kong, China
[4] Shenzhen Research Institute, Nanjing University of Aeronautics and Astronautics, Shenzhen, China

Abstract

Adverse weather conditions such as haze, rain, and snow often impair the quality of captured images, causing detection networks trained on normal images to generalize poorly in these scenarios. In this paper, we raise an intriguing question – if the combination of image restoration and object detection, can boost the performance of cutting-edge detectors in adverse weather conditions. To answer it, we propose an effective yet unified detection paradigm that bridges these two subtasks together via dynamic enhancement learning to discern objects in adverse weather conditions, called TogetherNet. Different from existing efforts that intuitively apply image dehazing/deraining as a pre-processing step, TogetherNet considers a multi-task joint learning problem. Following the joint learning scheme, clean features produced by the restoration network can be shared to learn better object detection in the detection network, thus helping TogetherNet enhance the detection capacity in adverse weather conditions. Besides the joint learning architecture, we design a new Dynamic Transformer Feature Enhancement module to improve the feature extraction and representation capabilities of TogetherNet. Extensive experiments on both synthetic and real-world datasets demonstrate that our TogetherNet outperforms the state-of-the-art detection approaches by a large margin both quantitatively and qualitatively. Source code is available at https://github.com/yz-wang/TogetherNet.

Keywords: TogetherNet, Object detection, Image restoration, Adverse weather, Joint learning, Dynamic transformer feature enhancement

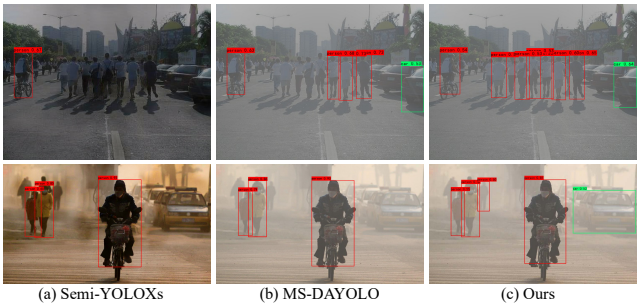
CCS Concepts

• **Computing methodologies** → *Object detection;*

1. Introduction

Object detection has been widely used in various practical real-world applications [CSKX15, HZL*21, BK21, WYB*22]. Despite the success of learning-based detectors on normal images, they usually fail to detect objects in images with adverse weather conditions, especially in hazy images [CLS*18, HLJ21]. This can be attributed to the noticeable degradation in image visibility and contrast caused by variant weather, which in turn drops the performance of object detectors. How to improve the accuracy of cutting-edging detectors in adverse weather conditions has attracted a great deal of attention [SOYP20, HR21, LRY*22, SCN21].

To tackle this challenging problem, an intuitive solution is to mitigate the effects of weather conditions by pre-processing the images using the restoration techniques such as image dehazing [LMSC19, LDR*20, DPX*20, RLHS20b, LLZX21]. Most of them



(a) Semi-YOLOXs (b) MS-DAYOLO (c) Ours

Figure 1: *Detection results by different methods on two typical examples of adverse weather conditions. From (a) to (c): the detection results by (a) Semi-YOLOXs [LDR*20] ("dehaze+detect"), (b) MS-DAYOLO [HR21], and (c) our TogetherNet.*

† Corresponding author (yxf@nuaa.edu.cn).

can enhance the overall visibility of these degraded images. But, can the restored images serve the downstream object detection task effectively? The answer may be not positive. That is because the restored images lose some important details which are beneficial to object detection [LRF*18]. *We consider the good way to serve object detection is to make object detection itself involve image restoration.* However, if simply combining a dehazing network with a detection network in a cascaded manner will increase the computational overhead and slow the inference time, which is undesirable in resource-constrained applications. *We further consider the good way to serve object detection in adverse conditions is to bridge image restoration and object detection together in a unified yet joint learning paradigm.*

In this paper, we respond to the intriguing learning-related question: combining a low-level image restoration task with a high-level object detection task to develop a multi-task joint learning paradigm will improve the performance of cutting-edge detection models in adverse weather conditions. Accordingly, we propose a novel unified paradigm that bridges these two subtasks together via dynamic enhancement learning for discerning objects in adverse weather conditions, termed a TogetherNet. Specifically, TogetherNet employs a cutting-edge object detector (i.e., YOLOX [GLW*21]) as the detection module, and exploits a feature restoration module to share the feature extraction module (backbone) with the detection network for image restoration. We train TogetherNet in an end-to-end fashion to simultaneously learn about image restoration and object detection. In this way, the latent information hidden in degraded images can be restored to benefit the detection task. In turn, the training of the detection task helps the backbone network to extract deeper structural and detailed features, thus facilitating the image restoration task. Moreover, considering that the performance of object detectors under adverse weather is usually limited, a Dynamic Transformer Feature Enhancement module (DTFE) is proposed to further enhance the feature extraction and representation capabilities of the model, thus improving its detection accuracy in such scenarios.

Recently, some approaches [SUHS19, LRY*22, HR21, RSA*21] cast object detection in adverse weather conditions as a task of learning models from a source domain (clean images) to a target domain (under adverse weather), i.e., unsupervised domain adaptation. These methods consider that compared with the clean images (source domain) used to train the detectors, the images (target domain) captured in adverse weather suffer from an obvious domain shift problem [GLC11, CLS*18]. They mostly employ domain adaptation strategies such as adversarial training to alien the target features with the source features. Despite the promising results achieved by domain adaptation under adverse weather, they usually ignore the latent information hidden in the degraded images which can also provide additional beneficial information for the detection task. As exhibited in Figure 1, compared with the "$dehaze + detect$", and domain adaptive-based detection models, the proposed TogetherNet can detect more objects with higher confidence, which demonstrates that our model outperforms the other algorithms for detecting objects in adverse weather. Note that Semi-YOLOXs is a typical "$dehaze + detect$" method that first restores the image and then detects the object, so the results in Figure 1 look different from other methods.

Extensive experiments on both synthetic benchmark (VOC-FOG-test) and real-world datasets (Foggy Driving Dataset [SDVG18] and RTTS [LRF*18]) demonstrate that our TogetherNet is far superior to the state-of-the-art object detection approaches. In summary, the main contributions are threefold:

- An effective yet unified detection paradigm is proposed for discerning objects in adverse weather conditions, which leverages a joint learning framework to perform image restoration and object detection tasks simultaneously, called, TogetherNet.
- We propose a Dynamic Transformer Feature Enhancement module (DTFE) to enhance the feature extraction and representation capabilities of TogetherNet.
- We compare TogetherNet with various representative state-of-the-art object detection approaches via extensive experiments, including "$dehaze + detect$", domain adaptive-based, multi-task-based, and image adaptive-based detection models. Consistently and substantially, TogetherNet performs favorably against them.

2. Related Work

In this section, we briefly summarize state-of-the-art object detectors that have produced encouraging results in general scenarios and under adverse weather conditions.

2.1. Object Detection

As a long-standing and fundamental task in computer vision, object detection has attracted extensive research attention in academia and industry [ZZXW19, LOW*20]. Recently, with the rapid development of convolutional neural networks (CNNs), learning-based detectors have dominated the modern object detection field for years.

Current object detection methods can be broadly categorized into two major groups, namely region proposal-based and regression-based. For region proposal-based approaches, they typically first employ methods such as selective search [UVDSGS13] to produce the candidate proposals, and then refine them for subsequent object detection. R-CNN [GDDM14] is the most representative region proposal-based detector, which adopts a CNN to extract features for the produced proposals, and then applies a support vector machine to perform classification. Inspired by the success of R-CNN in object detection, numerous variants based on this framework have sprung up, including Fast R-CNN [Gir15], Faster R-CNN [RHGS15], Libra R-CNN [PCS*19] and Dynamic R-CNN [ZCM*20]. Despite achieving encouraging detection accuracy, region proposal-based approaches are not satisfactory in terms of inference speed, which are undesirable in real-time applications. Therefore, to achieve a better speed-performance trade-off, various regression-based methods are developed for real-time detection. Representative approaches including YOLO series [RDGF16, RF17, RF18, BWL20, JSB21, GLW*21], SSD [LAE*16], RetinaNet [LGG*17], CenterNet [ZWK19], etc. In a nutshell, regression-based detectors are generally faster, but their detection performance is slightly weaker than region proposal-based detectors.

2.2. Object Detection in Adverse Weather

Compared with general object detection, few research efforts have been explored on object detection in adverse weather conditions.

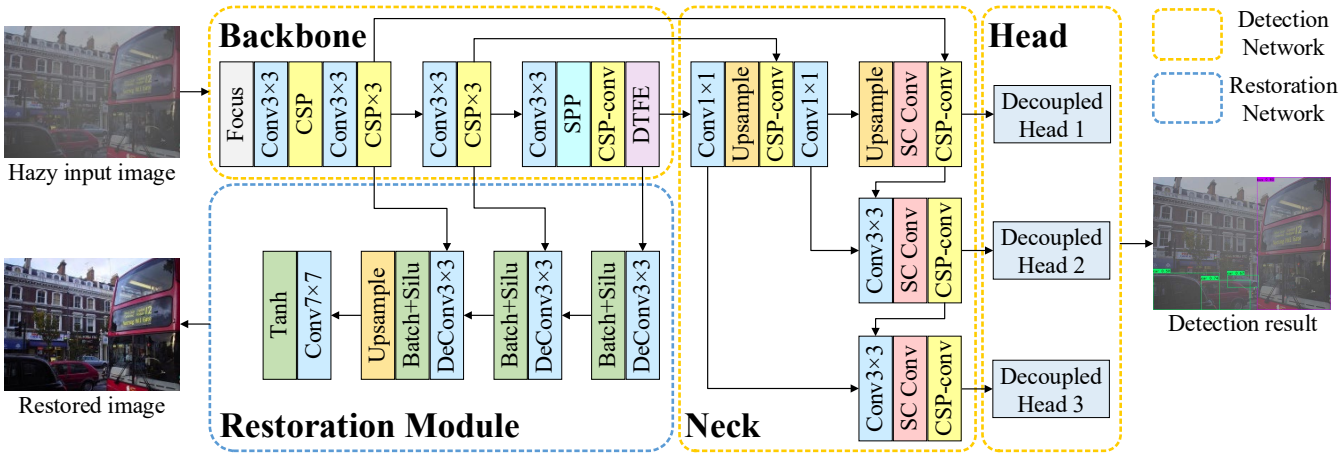


Figure 2: *The architecture of TogetherNet. It consists of both the object detection and image restoration networks. Note that the restoration module is only activated during the training phase. CSP/CSP-conv refers to Cross Stage Partial Network [WLW*20] with/without residual networks, DTFE refers to Dynamic Transformer Feature Enhancement module, and SC Conv refers to self-calibrated convolutions [LHC*20].*

Early methods mainly focused on pre-processing the degraded images by existing restoration algorithms such as image dehazing [HST11, QWB*20, WYG*22, SZB*22] or image deraining [LQS*19, RLHS20a, DWW*20], and then sending the processed images to the subsequent detection network for object detection. Although employing image restoration approaches as a pre-processing step can improve the overall quality of degraded images, these images may not definitely benefit the detection performance. A few prior-based efforts [LPW*17,HLJ21] have attempted to jointly perform image restoration and object detection to mitigate the effects of adverse weather-specific information. Sindagi et al. [SOYP20] develop a prior-based unsupervised domain adaptive framework for detecting objects in hazy and rainy conditions. Liu et al. [LRY*22] propose an image-adaptive detection network for object detection in adverse weather conditions, which combines image restoration and object detection into a unified framework and achieves very promising results. Recently, several methods [SLC19,ZPY*19,RCS*19,ZTHJ21] have begun to exploit domain adaption to overcome this problem. Zhang et al. [ZTHJ21] treat object detection in adverse weather as a domain shift problem and propose a domain adaptive YOLO to improve cross-domain performance for one-stage detectors.

3. TogetherNet

To boost the detection capacity of object detectors in adverse weather conditions, we come up with such a solution - is it possible to develop a unified detection paradigm that incorporates such detection task as a joint learning framework, while encouraging image restoration and detection tasks to benefit each other? If the answer is positive, our method can effectively address the detection of adverse weather scenarios. It is the focus of this work.

In this section, we first introduce the overview of the proposed unified detection paradigm, called TogetherNet, to demonstrate how we address the detection problem under adverse weather conditions. Next, the proposed restoration network is described in detail. After that, we elaborate on the proposed Dynamic Transformer Feature Enhancement module (DTFE) to explore the potential of deformable convolutions and self-attention mechanisms in feature extraction and representation. Finally, we describe the self-calibrated convolutions and Focal loss for optimizing our network to further improve the detection performance in adverse weather.

3.1. Overview of TogetherNet

The overall architecture of the proposed TogetherNet is depicted in Figure 2. Different from existing detection efforts, we consider overcoming the detection task from the following three perspectives. First, we employ an image restoration module to mitigate the influence of weather-specific information on the detection task. Second, a multi-task joint learning paradigm is developed to encourage low-level image restoration and high-level object detection tasks to collaborate and promote each other. Finally, a feature enhancement module is exploited to improve the feature extraction and representation capabilities of the model, such that more latent features can be revealed from degraded images to benefit image restoration and detection tasks.

YOLO series detectors [RDGF16, RF17, RF18, BWL20, JSB21, GLW*21] are the most representative regression-based detection models, which have been successfully applied in numerous scenarios. Recently, YOLOX [GLW*21] has been released as the latest version of the YOLO series detectors. Despite the promising results achieved by YOLOX in various benchmark datasets (e.g., MSCOCO [LMB*14], PASCAL-VOC [EVGW*10]), there are still many challenging yet unsolved problems. First, the YOLOX family detectors are originally designed for object detection in general yet easy scenes, without considering how to cope with the object detection in adverse weather conditions. Second, similar to most existing detectors, the YOLOX family detectors are susceptible to the weather-specific information in the detection task under adverse weather, resulting in a significant drop in detection accuracy. Third, YOLOXs (the smallest version of the YOLOX family)

is very lightweight and efficient, which is promising for resource-constrained mobile devices. But its detection performance is thus largely dropped for the adverse weather scenes.

To this end, we start from these three aspects and propose a novel unified detection paradigm for discerning objects in adverse weather conditions, called TogetherNet. To attain this objective, the proposed TogetherNet adopts one of the best object detectors, i.e., YOLOXs as our detection network to perform detection task. Our TogetherNet has the potential to benefit from a more complex version of the YOLOX family (e.g., YOLOXm and YOLOXl) to further improve its detection performance. However, we choose the smallest version of YOLOX because a lightweight model is more desirable in resource-limited/real-time applications.

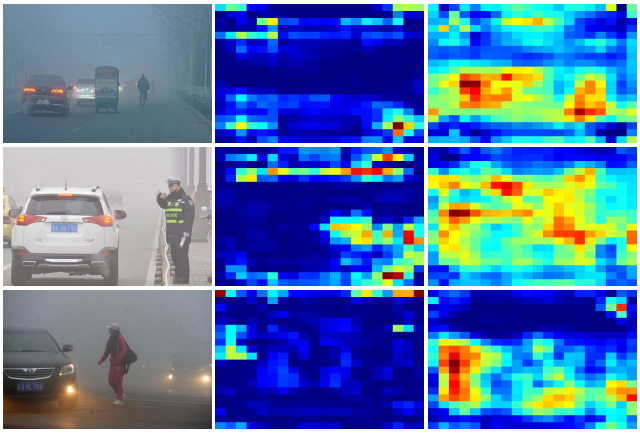
As depicted in Figure 2, the proposed TogetherNet consists of two main modules, i.e., the detection network and the restoration network. Given a hazy input image, we first employ the focus operation in the backbone module to separate the image into different granularities and regroup them together to enhance the image features. Then, several cross stage partial modules (CSP) [WLW*20] and a Dynamic Transformer Feature Enhancement module (DTFE) are employed to extract the complex and latent features from the restructured feature map. DTFE is a novel feature enhancement module developed to expand the receptive field with adaptive shape and enhance the model's feature representation capability for better detection and image restoration. After that, the extracted features are transmitted to both the restoration module and the neck module to perform different tasks. In this way, TogetherNet can benefit from the joint learning framework, where the clean features produced by the restoration module can be shared to learn better object detection in the detection network. Finally, the detection head module will produce the final class probability scores, bounding boxes, and confidence scores.

Moreover, to further improve the detection capacity of TogetherNet and well address the challenge of detecting objects in adverse weather, we introduce a multi-scale feature enhancement module, namely, self-calibrated convolutions and the well-known Focal loss into our model. Both of them have been widely used in object detection networks and proved to be effective in improving detection accuracy, which will be described in the following sections. We emphasize that, since this work mainly focuses on object detection in adverse weather, and introducing the image restoration module in the testing phase would significantly slow down TogetherNet's inference speed, the restoration module is only activated during the training phase.

3.2. Restoration Network

In our design, the restoration network is responsible for recovering the clean images and sharing these restored features with the detection network during joint learning to promote the model's detection accuracy in adverse weather conditions. To attain this objective, we employ the backbone network to extract the complex and latent features hidden in the input image for simultaneously learning image restoration and object detection.

Considering that the features extracted by the backbone network may be degraded by weather-specific information (i.e. haze), resulting in poor detection performance, a decoder-like network is developed as our restoration module to eliminate these effects as much as possible. As demonstrated in Figure 2, to restore the clean image features, three deconvnet, an up-sampling operation, and a *Tanh* activation function are adopted to produce the final clean images. Moreover, we also introduce the skip connection strategy to facilitate the detection task by revealing multi-scale latent features while avoiding the gradient vanishing problem.



(a) Input (b) Ours w/o restoration loss (c) Ours

Figure 3: *Visualization of feature maps.*

To perform image restoration, the mean square error (MSE) loss is employed to train the restoration network, which can be expressed by:

$$L_{re} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mathbf{Y}}_i - \mathbf{Y}_i\right)^2, \tag{1}$$

where n denotes the batch size, $\hat{\mathbf{Y}}_i$ refers to the ground truth image, and $\mathbf{Y}_i$ refers to the estimated clean image. Actually, using a more complex network architecture or loss function may enhance the dehazing performance of current models, we prefer to adopt a simple CNN-based network and MSE loss to achieve a better parameter-performance trade-off.

To better understand the effectiveness of the proposed restoration network, we visualize the features of the last layer in our backbone module (with/without restoration loss). As depicted in Figure 3, features with restoration loss mitigate the influence of weather information on them to some extent, and can still focus on regions containing objects (see the red regions in Figure 3), allowing for better performing detection tasks.

Furthermore, we have tried to send the clean images recovered by the restoration network directly to the detection module for object detection, but the detection accuracy appears to be greatly dropped. We argue that the restored clean image weakens some features of the original image, and even creates a new domain shift problem during the image restoration, prohibiting such a strategy from achieving optimal performance. In light of this, we consider employing the restoration network to produce the latent clean

features from the backbone network through learning the image restoration task. As a consequence, the improved detection performance of TogetherNet can be achieved by jointly optimizing image restoration and object detection.

3.3. Dynamic Transformer Feature Enhancement Module

Fundamentally, the feature extraction and representation capabilities of the network directly determine the performance of the model. Hence, we argue that there are two solutions to reduce the impact of weather-degradations on detection tasks under adverse weather. The first approach is to expand the receptive field of the network to help the model fuse more spatially structured information, so that the objects can be discerned from the area less affected by weather-specific information. Another approach is to enhance the feature extraction capability of the network so that objects can be detected directly from these areas with poor visibility. To this end, we develop a novel Dynamic Transformer Feature Enhancement module (DTFE) to improve the model's feature extraction and representation capabilities for better image restoration and object detection. The DTFE module mainly consists of two parts, i.e., a dynamic feature transformation network (DFT) and a Transformer-based feature enhancement network (TFE), as depicted in Figure 4. Specifically, we employ two deformable convolutions [DQX*17] to form the dynamic feature transformation network, which can expand the receptive field of the model with adaptive shape and improve its transformation ability. For the feature enhancement network, we adopt the Vision Transformer block [DBK*20] to explore the potential of the self-attention mechanisms in improving the model's feature representation capability.

Different from conventional CNNs, the kernels in deformable convolutions are dynamic and flexible, which can capture more spatially structured information. Moreover, the work [ZSL*20] has demonstrated that dynamic and flexible convolution kernels can effectively improve the feature transformation capabilities of networks, such that the deformable convolutions are employed to enhance the feature for object detection. Therefore, we develop a dynamic feature transformation network based on deformable convolutions to expand the receptive field with adaptive shape and improve the model's transformation capability (see Figure 4). In this way, our model can focus on more areas that are less affected by weather-specific information, thus reducing the impact of weather-degradations on detection accuracy.

Recently, Vision Transformers (ViTs) have become one of the dominant models in computer vision owing to their ability to learn complex dependencies between input features via self-attention mechanisms. Given this, we consider adopting the ViT module in the design of the backbone network to boost its feature representation ability, thus improving the performance of the subsequent image restoration and object detection tasks. In particular, we introduce a feature enhancement network via the Vit module in the last layer of the backbone network to further enhance the extracted features. It enables the backbone network to build complex and long-range spatial dependencies between the input features, thus improving the detection capacity of our TogetherNet.

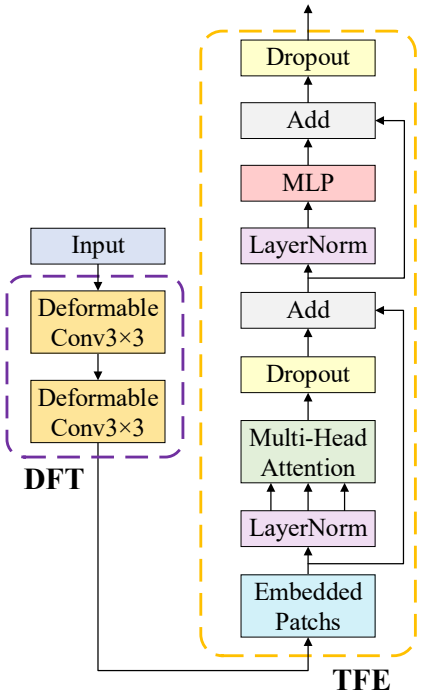


Figure 4: *The architecture of the Dynamic Transformer Feature Enhancement module (DTFE). DTFE can help the backbone network improve its feature extraction and representation capabilities for better image restoration and object detection. DFT refers to dynamic feature transformation network, TFE refers to Transformer-based feature enhancement network.*

3.4. Self-calibrated Convolutions

The self-calibrated convolution network is an improved CNN structure proposed by Liu et al. [LHC*20], which can build long-range spatial and inter-channel dependencies around each spatial location. Therefore, it can enlarge the receptive field of each convolutional layer and enhance the feature extraction ability of CNNs. In light of this, we consider adopting the self-calibrated convolution network as a multi-scale feature extraction module to cope with the weather-degradation problem in the detection task and improve the detection performance of TogetherNet.

The architecture of the self-calibrated convolution network is exhibited in Figure 5. Given an input feature map X with channel C, we first split it into two feature maps X_1 and X_2 with channel $C/2$. Then, we send X_1 to the self-calibrated branch for feature transformation and fusion. In this branch, three filters (K_2, K_3, and K_4) are employed to extract and fuse multi-scale features from X_1. Next, we employ a filter K_1 to transmit and extract features from X_2 to obtain the other half of the result Y_2. Finally, Y_1 and Y_2 are concatenated to produce the final output Y. In our designation, we introduce self-calibrated convolutions in front of the three decoupled YOLO heads to expand the receptive field of the convolutional layer and extract multi-scale features for better object detection (see Figure 2). In this way, our TogetherNet can well address the challenge of discerning objects in adverse weather conditions.

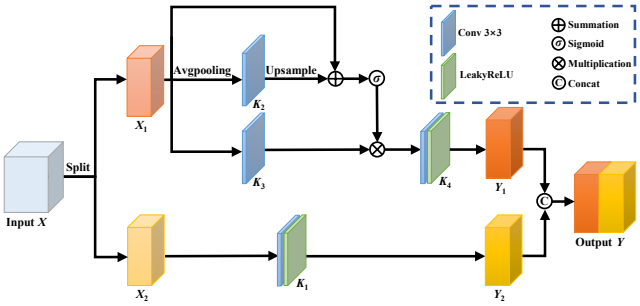


Figure 5: *The architecture of the self-calibrated convolutions. It can build long-range spatial and inter-channel dependencies around each spatial location, thus expanding the receptive field of each convolutional layer and enhancing the feature extraction capability of CNNs.*

Moreover, inspired by RetinaNet [LGG*17] with high effective Focal loss, we introduce the Focal loss in the design of our TogetherNet to address the problem of imbalance between positive and negative samples in object detection tasks. Therefore, the total loss function can be formulated as:

$$L_{Total} = L_{de} + L_{re}, \tag{2}$$

where L_{de} refers to the detection loss, which can be expressed by:

$$L_{de} = \lambda L_{IoU} + L_{Cls} + L_{Focal}, \tag{3}$$

where λ is loss weight, and we set $\lambda = 5$ here. L_{IoU} and L_{Cls} refer to regression loss and classification loss, respectively.

In our experiments, we are surprised to find that the image restoration loss L_{re} is very helpful in improving the performance of the detection task, here we set the loss weight of L_{re} to 0.8, and the loss weight of L_{de} to 0.2. For these two loss weights, extensive experiments are performed to ensure their optimum values (see Section 4.5). Therefore, developing a joint learning paradigm that combines these two tasks is very effective to improve detection capacity in adverse weather conditions.

4. Experiments

In this section, comprehensive experiments are performed to evaluate the detection performance of TogetherNet and other detection approaches under adverse weather conditions. To conduct experiments, a dataset for detecting objects in foggy weather is established, called VOC-FOG. For evaluation, Both the synthetic foggy dataset (VOC-FOG-test) and real-world foggy datasets (Foggy Driving dataset [SDVG18] and Real-world Task-driven Testing Set (RTTS) [LRF*18]) are employed as the testing set. All the experiments are implemented by PyTorch 1.9 on a system with an Intel(R) Core(TM) i7-9700 CPU, 16 GB RAM, and an NVIDIA GeForce RTX 3090 GPU.

4.1. Dataset

Considering there are few publicly available datasets for object detection in adverse weather conditions, to train and evaluate the proposed TogetherNet, we establish a foggy detection dataset based on the classic VOC dataset [EVGW*10], dubbed VOC-FOG. Specifically, we employ the well-known atmospheric scattering model to generate the foggy images $I(x)$, which can be obtained by the following formula:

$$I(x) = J(x)t(x) + A(1 - t(x)), \tag{4}$$

where $J(x)$ denotes the clean image, A refers to the global atmospheric light, and $t(x)$ refers to the medium transmission map, which can be calculated by:

$$t(x) = e^{-\beta d(x)}, \tag{5}$$

where β denotes the atmosphere scattering parameter, and $d(x)$ refers to the scene depth, which can be defined as:

$$d(x) = -0.04 \times \rho + \sqrt{max(w,h)}, \tag{6}$$

where ρ denotes the Euclidean distance from the current pixel to the central pixel, w and h refer to the numbers of rows and columns of the image. In our experiments, we set the global atmospheric light parameter A to 0.5, while randomly setting the atmospheric scattering parameter β between 0.07 and 0.12 to control the fog level. Moreover, considering there are five annotated object classes (i.e., car, bus, motorcycle, bicycle, and person) in the RTTS dataset, to form our training dataset, we select the images containing these five categories to add haze. After processing these clean images on the original VOC dataset, we obtain 9578 foggy images for training (VOC-FOG) and 2129 images for testing (VOC-FOG-test), as depicted in Table 1 and Figure 6.

As observed, the fog in the central area appears to be thicker than in the surrounding areas, which can be explained by the principle of synthetic fog. When employing the atmospheric scattering model to generate fog, we first need to select a point as the starting point and

Table 1: *Details about training and testing dataset. The object classes are bic (bicycle), bus, car, mot (motorcycle), and per (person). FDD is the abbreviation of Foggy Driving Dataset.*

Dataset	Images	Bic	Bus	Car	Mot	Per
VOC-FOG	9578	836	684	2453	801	13519
VOC-FOG-test	2129	155	156	857	131	3527
FDD	101	17	17	425	9	269
RTTS	4332	534	1838	18413	862	7950

(a) Clean images

(b) Synthetic foggy images

Figure 6: *Example images in the proposed VOC-FOG dataset.*

Table 2: *Comparison of TogetherNet with state-of-the-art detection models on the VOC-FOG-test dataset. * denotes that the model is trained with clean images from the VOC-FOG dataset. Red and blue colors are used to indicate the 1^{st} and 2^{nd} ranks, respectively.*

Method	Publication	Type	Person	Bicycle	Car	Motorbike	Bus	*mAP*
YOLOXs [GLW*21]	arXiv'21	Baseline	80.81	74.14	83.63	75.35	86.40	80.07
YOLOXs* [GLW*21]	arXiv'21	Baseline	79.97	67.95	74.75	58.62	83.12	72.88
DCP-YOLOXs* [HST11]	TPAMI'11	Dehaze	81.58	78.80	79.75	78.51	85.64	80.86
AOD-YOLOXs* [LPW*17]	ICCV'17	Dehaze	81.26	73.56	76.98	71.18	83.08	77.21
Semi-YOLOXs* [LDR*20]	TIP'20	Dehaze	81.15	76.94	76.92	72.89	84.88	78.56
FFA-YOLOXs* [QWB*20]	AAAI'20	Dehaze	78.30	70.31	69.97	68.80	80.72	73.62
MS-DAYOLO [HR21]	ICIP'21	Domain adaptive	82.52	75.62	86.93	81.92	90.10	83.42
DS-Net [HLJ21]	TPAMI'21	Multi-task	72.44	60.47	81.27	53.85	61.43	65.89
IA-YOLO [LRY*22]	AAAI'22	Image adaptive	70.98	61.98	70.98	57.93	61.98	64.77
TogetherNet	ours	Multi-task	87.62	78.19	85.92	84.03	93.75	85.90

then spread the synthetic fog around it. Considering that the center of natural images is generally the position with the largest depth value, it is common to use the center point as the starting point when synthesizing fog. Therefore, the fog in the center is usually thicker.

Testing set. To evaluate the detection performance of TogetherNet and other detection methods in adverse weather conditions, both the synthetic foggy dataset (VOC-FOG-test) and two real-world foggy datasets (Foggy Driving dataset and RTTS) are employed as our testing set.

- **VOC-FOG-test** contains 2129 foggy images synthesized from the clean images in the VOC dataset. Different from the above-mentioned VOC-FOG training set, to further verify the generalization ability of TogetherNet, we set atmosphere scattering parameter β to a wider range to simulate extreme foggy and misty weather conditions. Specifically, the value of β is randomly set between 0.05 and 0.14 to adjust for different fog levels.
- **Foggy Driving Dataset** [SDVG18] is a real-world foggy dataset that is used for object detection and semantic segmentation. It involves 466 vehicle instances (i.e., car, bus, train, truck, bicycle, and motorcycle) and 269 human instances (i.e., person and rider) that are labeled from 101 real-world foggy images. Furthermore, although there are eight annotated object classes in the Foggy Driving Dataset, we only select the above-mentioned five object classes for detection to ensure consistency between training and testing.
- **RTTS** [LRF*18] is a relatively comprehensive dataset available in natural foggy conditions, which comprises 4322 real-world foggy images with five annotated object classes. Considering hazy/clean image pairs are difficult or even impossible to capture in the real world, Li et al. proposed the RTTS dataset to evaluate the generalization ability of dehazing algorithms in real-world scenarios from a task-driven perspective.

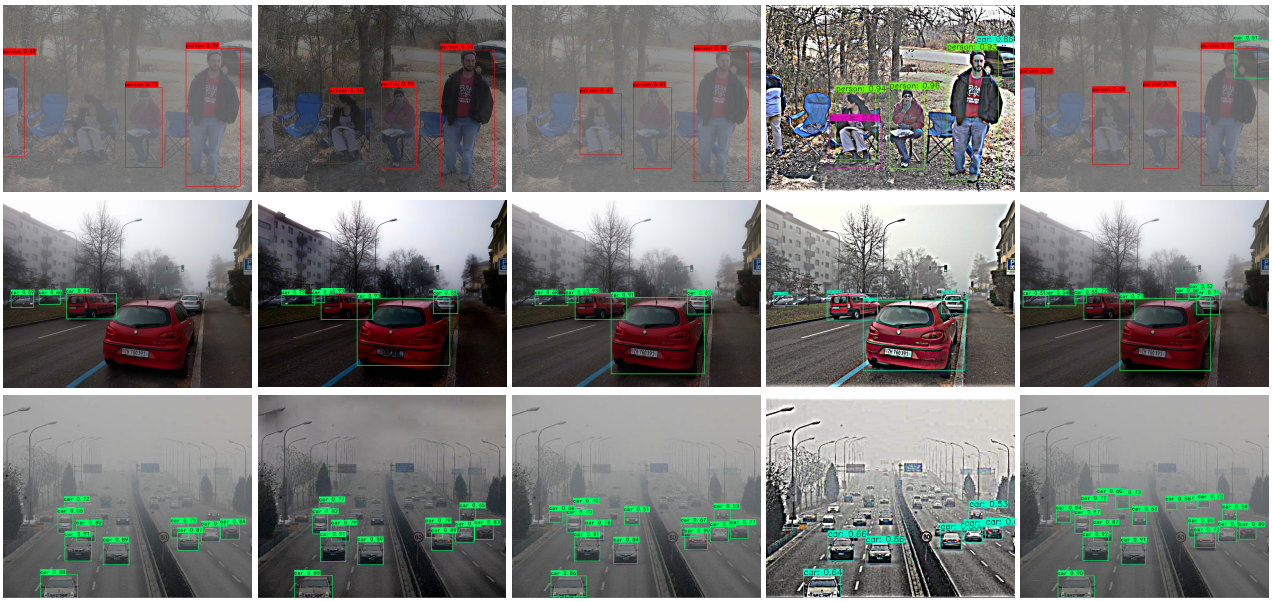
4.2. Implementation Details

Training details. TogetherNet is trained using the SGD optimizer with a batch size of 16. The initial learning rate l is set to 1×10^{-2}. We empirically set the total number of epochs to 100 and adopt a Cosine annealing decay strategy to adjust the learning rate l. In addition to feeding foggy images to TogetherNet for training the detection task, we also send the original clean images from the VOC-FOG dataset for training the image restoration task. Both the training and testing images are resized to 640×640. Moreover, we did not add the mosaic data augmentation strategy in the training process as YOLOX defaults, because adopting this approach would increase the difficulty of training our image restoration network, which in turn drops the performance of the object detection task.

Evaluation Settings. To quantitatively evaluate the performance of the proposed TogetherNet, we adopt the mean Average Precision (*mAP*) as the evaluation metric, which is the most widely used objective evaluation index in object detection tasks. We compare TogetherNet with various state-of-the-art object detection approaches. These object detection methods can be classified into four categories: 1) "$dehaze + detect$" methods: Here, we employ several dehazing algorithms as a pre-processing step and perform object detection by YOLOXs trained on clean VOC images (the original clean images from VOC-FOG dataset). For pre-processing, we chose four popular dehazing approaches, namely, DCP [HST11], AOD-Net [LPW*17], Semi-dehazing [LDR*20], and FFA-Net [QWB*20] to combine with the YOLOXs detector for forming four combination models called DCP-YOLOXs, AOD-YOLOXs, Semi-YOLOXs, and FFA-YOLOXs, respectively; 2) domain-adaptive-based MS-DAYOLO [HR21]; 3) multi-task-based DS-Net [HLJ21]; and 4) image adaptive-based IA-YOLO [LRY*22]. Note that all the dehazing algorithms are trained on the entire ITS (Indoor Training Set) [LRF*18] dataset according to the settings in their papers.

4.3. Comparison with State-of-the-arts

Comparison on Synthetic Dataset. The *mAP* metric of ten detection algorithms on the proposed VOC-FOG-test dataset are reported for quantitative evaluation, as demonstrated in Table 2. To make a fair comparison, we retrain all compared methods (except "$dehaze + detect$" methods) on the proposed VOC-FOG dataset according to the settings in their papers. For "$dehaze + detect$" methods, we found that if the baseline YOLOXs is trained on the foggy dataset (VOC-FOG), the final detection results on the testing set will be dropped no matter what dehazing algorithm is employed. This could be an obvious domain shift between the training set (hazy images) and testing set (dehazed images), resulting in a significant decrease in detection accuracy. Therefore, we



(a)YOLOXs (b) Semi-YOLOXs (c) MS-DAYOLO (d) IA-YOLO (e) Ours

Figure 7: *Detection results by different methods on both synthetic and real-world foggy datasets. From (a) to (e): the detection results by (a) YOLOXs [GLW*21], (b) Semi-YOLOXs [LDR*20], (c) MS-DAYOLO [HR21], (d) IA-YOLO [LRY*22], and (e) our TogetherNet, respectively. Clearly, the proposed TogetherNet can discern more objects with higher confidence.*

Table 3: *Quantitative comparisons (mAP) with state-of-the-art detection approaches on the Foggy Driving Dataset.*

Method	Person	Bicycle	Car	Motorcycle	Bus	*mAP*
YOLOXs	23.26	26.55	55.04	7.14	45.71	31.54
YOLOXs*	26.69	23.04	56.27	2.38	41.98	30.07
DCP-YOLOXs*	22.24	10.78	56.34	7.14	50.66	29.43
AOD-YOLOXs*	24.54	33.82	56.75	4.76	36.04	31.18
Semi-YOLOXs*	22.39	27.73	56.47	4.76	44.93	31.26
FFA-YOLOXs*	19.18	18.07	50.83	2.38	42.77	26.65
MS-DAYOLO	21.52	34.58	56.39	8.33	47.68	33.70
DS-Net	26.74	20.54	58.16	7.14	36.11	29.74
IA-YOLO	16.20	11.76	41.43	4.76	17.55	18.34
TogetherNet	30.44	28.44	58.24	14.29	43.23	34.93

Table 4: *Quantitative mAP values of the proposed TogetherNet and various state-of-the-art detection approaches on the RTTS dataset. Clearly, our TogetherNet achieves the best performance.*

Method	Person	Bicycle	Car	Motorbike	Bus	*mAP*
YOLOXs	77.23	40.55	68.72	40.83	28.81	51.23
YOLOXs*	76.07	48.47	63.88	41.03	22.76	50.44
DCP-YOLOXs*	76.81	50.03	62.84	40.62	23.73	50.81
AOD-YOLOXs*	76.49	43.32	61.03	34.54	22.16	47.51
Semi-YOLOXs*	75.71	46.72	62.74	40.37	24.51	50.01
FFA-YOLOXs*	76.52	48.13	64.31	39.74	23.71	50.48
MS-DAYOLO	74.22	44.11	69.73	37.54	36.45	52.41
DS-Net	68.81	18.02	46.13	15.15	15.44	32.71
IA-YOLO	67.25	35.28	41.14	20.97	13.64	35.66
TogetherNet	82.70	57.27	75.32	55.40	37.04	61.55

adopted clean images from the VOC-FOG dataset to train the baseline YOLOXs for these methods. As observed, our TogetherNet outperforms other state-of-the-arts by a large margin in accuracy rate.

Comparison on Real-world Dataset. We also compare our TogetherNet with several state-of-the-art methods on two real-world foggy datasets, namely, Foggy Driving Dataset and RTTS dataset. Table 3 and Table 4 exhibit the *mAP* metric of all compared methods on these two real-world foggy datasets. Different from the results in the VOC-FOG-test dataset, the "$dehaze + detect$" methods are very limited in improving detection accuracy in real-world degraded scenarios, validating that these processed images do not always guarantee improved object detection performance. clearly, our TogetherNet achieves the highest *mAP* values again on both datasets, compared to the SOTAs.

For qualitative comparisons, we exhibit three detection results from the VOC-FOG-test, Foggy Driving, and RTTS datasets in Figure 7. Our TogetherNet is compared with YOLOXs baseline, Semi-YOLOXs, MS-DAYOLO, and IA-YOLO. As observed, TogetherNet can detect more objects with higher confidence, which demonstrates that our approach performs well in both synthetic and real-world foggy datasets. Similar to Semi-YOLOXs, IA-YOLO is also a paradigm of first enhancing the image and then detecting the object, thus they look different from the other approaches.

4.4. Experiments on Rainy Images

To demonstrate that the proposed model can generalize well under other adverse weather conditions, we adopt the RainCityscapes

Table 5: *Comparison of TogetherNet with baseline YOLOXs and Syn-YOLOXs ("derain+detect") methods on the RainCityscapes dataset. * denotes that the model is trained with clean images from the RainCityscapes dataset.*

Method	Person	Bike	Car	Motorbike	Bus	*mAP*
YOLOXs	21.89	30.37	52.36	1.26	21.15	25.41
YOLOXs*	23.04	15.14	64.47	0.03	13.61	23.26
Syn-YOLOXs*	17.11	13.29	62.22	2.03	24.55	23.84
TogetherNet	19.64	19.98	64.38	12.94	25.28	28.44

(a) YOLOXs (b) Syn-YOLOXs (c) Ours

Figure 8: *Detection results by different methods on the RainCityscapes dataset. From (a) to (c): the detection results by (a) YOLOXs [GLW*21], (b) Syn-YOLOXs [YSP20] ("derain+detect"), and (c) our TogetherNet.*

dataset [HZW*21] to evaluate the detection performance of TogetherNet in rainy weather. RainCityscapes dataset contains 10,620 synthetic rainy images (295 images with 36 variants) with eight annotated object classes: car, train, truck, motorbike, bus, bike, rider, and person. In our experiments, we randomly choose 2,500 rainy images (250 images with 10 variants) for training and 450 images (45 images with 10 variants) for testing. As in the previous experimental setup, we only select the aforementioned five object categories for detection.

We compare our TogetherNet with the baseline YOLOXs and a "$derain + detect$" (Syn2Real [YSP20]) method on the testing set. The *mAP* metric of these 3 detection algorithms is reported in Table 5. As can be seen, our TogetherNet outperforms other approaches by a large margin again in accuracy rate. Figure 8 exhibits three visual examples of the baseline YOLOXs, the "$derain + detect$" method, and our TogetherNet. As observed, the proposed TogetherNet can discern more objects with higher confidence, which demonstrates that our approach generalizes well under rainy weather conditions.

4.5. Ablation Study

Effect of different components in TogetherNet. The proposed network exhibits superior detection performance compared to the state-of-the-art detection methods. To further evaluate the effectiveness of TogetherNet, we conduct extensive ablation studies to analyze the different components, including the image restoration module, Dynamic Transformer Feature Enhancement module (DTFE), Focal loss, and self-calibrated convolutions.

Table 6: *Ablation study of different training strategies on the RTTS dataset. Clearly, our full model (V_4) outperforms other alternatives. IR is the abbreviation of the image restoration module.*

Variants	Base	V_1	V_2	V_3	V_4	V_5	V_6	V_7
IR	w/o	✓	✓	✓	✓	w/o	✓	✓
DTFE	w/o	w/o	✓	✓	✓	✓	w/o	✓
Focal loss	w/o	w/o	w/o	✓	✓	✓	✓	w/o
SC Conv	w/o	w/o	w/o	w/o	✓	✓	✓	✓
mAP	51.23	56.05	57.73	59.83	**61.55**	54.97	56.57	56.75

Table 7: *Ablation study on the object detection loss L_{de} and image restoration loss L_{re} (loss weight λ_1 and λ_2).*

λ_1&λ_2	1&1	0.7&0.3	0.5&0.5	0.2&0.6	0.2&0.8	0.2&1.0	0.1&1.2
mAP	53.79	57.44	57.73	58.30	**61.55**	60.08	58.09

We first construct our base network with the original YOLOXs detector as the baseline of the detection network, and then we train this model with the implementation details mentioned above. Next, we incrementally add different components into the base network as follows:

1. base model + image restoration module $\rightarrow V_1$,
2. V_1 + DTFE $\rightarrow V_2$,
3. V_2 + Focal loss $\rightarrow V_3$,
4. V_3 + self-calibrated convolutions $\rightarrow V_4$ (full model),
5. V_4 - image restoration module $\rightarrow V_5$,
6. V_4 - DTFE $\rightarrow V_6$,
7. V_4 - Focal loss $\rightarrow V_7$,

All these variants are retrained in the same way as before and tested on the RTTS dataset. The performances of these models are depicted in Table 6.

As observed, each component in TogetherNet helps in improving object detection performance, especially the proposed image restoration module, which achieves 4.82 *mAP* gains over our base model. The introduction of the proposed DTFE, Focal loss, and self-calibrated convolutions also greatly improved the performance of the model. In short, if we make full use of the implementation details in this paper, the detection results will outperform other competitive approaches.

Effect of the weights in loss functions. To improve the detection performance of TogetherNet in adverse weather conditions, we exploit an effective unified loss function that contains object detection loss L_{de} and image restoration loss L_{re}. Accordingly, two loss weights (λ_1 and λ_2) are employed to balance the performance of these two loss functions. For λ_1 and λ_2, extensive experiments are conducted on the RTTS dataset to ensure their optimum values, as exhibited in Table 7. As observed, the image restoration loss is very helpful in improving the detection capacity of the proposed method. Therefore, when setting $\lambda_1 = 0.2$ and $\lambda_2 = 0.8$ in our experiments, the performance of TogetherNet is the best.

(a)YOLOXs (b) Semi-YOLOXs (c) MS-DAYOLO (d) IA-YOLO (e) Ours

Figure 9: *Typical failure cases of different detection algorithms. None of the detectors can discern all the objects in such images.*

Table 8: *Runtime (in seconds) and FPS comparisons of different detection methods tested on an image of 550 × 400 pixels. * indicates the platform used for image dehazing.*

Method	Platform	Run time	*FPS*
YOLOXs [GLW*21]	PyTorch (GPU)	0.018	55.6
DCP-YOLOXs [HST11]	Python (CPU)*	1.238	0.8
AOD-YOLOXs [LPW*17]	PyTorch (GPU)*	0.121	8.3
Semi-YOLOXs [LDR*20]	PyTorch (GPU)*	1.108	0.9
FFA-YOLOXs [QWB*20]	PyTorch (GPU)*	0.366	2.7
MS-DAYOLO [HR21]	Caffe (GPU)	0.037	27.0
DS-Net [HLJ21]	PyTorch (GPU)	0.035	28.6
IA-YOLO [LRY*22]	Tensorflow (GPU)	0.039	25.6
TogetherNet (ours)	PyTorch (GPU)	0.031	32.3

4.6. Efficiency Analysis

Considering efficiency is essential for a computer vision system, we evaluate the computational performance of various state-of-the-art detection methods and report their average running times and frames per second (*FPS*) metrics in Table 8. All the approaches are implemented on a system with an Intel(R) Core(TM) i7-9700 CPU, 16 GB RAM, and an NVIDIA GeForce RTX 3090 GPU. It can be seen that our TogetherNet takes about 0.031*s* to infer an image of 550 × 400 pixels on average. The proposed TogetherNet is fast and efficient since it ranks second among the ten detection algorithms.

4.7. Limitation and Discussion

Although TogetherNet has achieved encouraging results on both synthetic and real-world foggy datasets, our model is not very robust for the heavily foggy scene. We provide two typical failure cases in Figure 9. It can be observed that the heavy fog degrades the performance of various object detectors. Even humans have difficulty discerning the objects in such challenging images. This limitation might be solved by introducing more effective feature enhancement modules in our network. In near future, we will make efforts to solve this limitation.

5. Conclusion

We propose an efficient unified detection paradigm for discerning objects in adverse weather conditions, named TogetherNet. It leverages a joint learning framework to perform image restoration and object detection tasks simultaneously. From a different yet new perspective, TogetherNet casts such detection task as multi-task joint learning, where these two tasks are collaborated and contributed to each other. To better cope with the weather-degradations in this detection task, we develop a Dynamic Transformer Feature Enhancement module (DTFE) to enhance the feature extraction and representation capabilities of our model. In addition, the self-calibrated convolution network is introduced to expand the receptive field of each convolutional layer and enrich the output features, thus reducing the impact of weather-specific information on detection accuracy. Furthermore, we also employ the well-known Focal loss to address the problem of imbalance between positive/negative samples in detection tasks. Experiments on both synthetic and real-world foggy datasets demonstrate that our TogetherNet performs favorably against state-of-the-art detection algorithms.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (No. 62172218), the Joint Fund of National Natural Science Foundation of China and Civil Aviation Administration of China (No. U2033202), the 14th Five-Year Planning Equipment Pre-Research Program (No. JCKY2020605C003), the Free Exploration of Basic Research Project, Local Science and Technology Development Fund Guided by the Central Government of China (No. 2021Szvup060), the Natural Science Foundation of Guangdong Province (No. 2022A1515010170).

References

[BK21] BOZCAN I., KAYACAN E.: Context-dependent anomaly detection for low altitude traffic surveillance. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)* (2021), pp. 224–230. 1

[BWL20] BOCHKOVSKIY A., WANG C.-Y., LIAO H.-Y. M.: Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934* (2020). 2, 3

[CLS*18] CHEN Y., LI W., SAKARIDIS C., DAI D., VAN GOOL L.: Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2018), pp. 3339–3348. 1, 2

[CSKX15] CHEN C., SEFF A., KORNHAUSER A., XIAO J.: Deepdriving: Learning affordance for direct perception in autonomous driving. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (2015), pp. 2722–2730. 1

[DBK*20] DOSOVITSKIY A., BEYER L., KOLESNIKOV A., WEISSENBORN D., ZHAI X., UNTERTHINER T., DEHGHANI M., MINDERER M., HEIGOLD G., GELLY S., ET AL.: An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020). 5

[DPX*20] DONG H., PAN J., XIANG L., HU Z., ZHANG X., WANG F., YANG M.-H.: Multi-scale boosted dehazing network with dense feature fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), pp. 2157–2167. 1

[DQX*17] DAI J., QI H., XIONG Y., LI Y., ZHANG G., HU H., WEI Y.: Deformable convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (2017), pp. 764–773. 5

[DWW*20] DENG S., WEI M., WANG J., FENG Y., LIANG L., XIE H., WANG F. L., WANG M.: Detail-recovery image deraining via context aggregation networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), pp. 14548–14557. 3

[EVGW*10] EVERINGHAM M., VAN GOOL L., WILLIAMS C. K., WINN J., ZISSERMAN A.: The pascal visual object classes (voc) challenge. *International Journal of Computer Vision 88*, 2 (2010), 303–338. 3, 6

[GDDM14] GIRSHICK R., DONAHUE J., DARRELL T., MALIK J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2014), pp. 580–587. 2

[Gir15] GIRSHICK R.: Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (2015), pp. 1440–1448. 2

[GLC11] GOPALAN R., LI R., CHELLAPPA R.: Domain adaptation for object recognition: An unsupervised approach. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (2011), IEEE, pp. 999–1006. 2

[GLW*21] GE Z., LIU S., WANG F., LI Z., SUN J.: Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430* (2021). 2, 3, 7, 8, 9, 10

[HLJ21] HUANG S.-C., LE T.-H., JAW D.-W.: Dsnet: Joint semantic learning for object detection in inclement weather conditions. *IEEE Transactions on Pattern Analysis and Machine Intelligence 43*, 8 (2021), 2623–2633. 1, 3, 7, 10

[HR21] HNEWA M., RADHA H.: Multiscale domain adaptive yolo for cross-domain object detection. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)* (2021), IEEE, pp. 3323–3327. 1, 2, 7, 8, 10

[HST11] HE K., SUN J., TANG X.: Single image haze removal using dark channel prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence 33*, 12 (2011), 2341–2353. 3, 7, 10

[HZL*21] HE L., ZHOU Q., LI X., NIU L., CHENG G., LI X., LIU W., TONG Y., MA L., ZHANG L.: End-to-end video object detection with spatial-temporal transformers. In *Proceedings of the 29th ACM International Conference on Multimedia* (2021), pp. 1507–1516. 1

[HZW*21] HU X., ZHU L., WANG T., FU C.-W., HENG P.-A.: Single-image real-time rain removal based on depth-guided non-local features. *IEEE Transactions on Image Processing 30* (2021), 1759–1770. 9

[JSB21] JOCHER G., STOKEN A., BOROVEC J.: Ultralytics/yolov5: V4.0–nn.silu() activations, weights & biases logging, pytorch hub integration, Jan, 2021. URL: https://zenodo.org/record/4418161. 2, 3

[LAE*16] LIU W., ANGUELOV D., ERHAN D., SZEGEDY C., REED S., FU C.-Y., BERG A. C.: Ssd: Single shot multibox detector. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2016), Springer, pp. 21–37. 2

[LDR*20] LI L., DONG Y., REN W., PAN J., GAO C., SANG N., YANG M.-H.: Semi-supervised image dehazing. *IEEE Transactions on Image Processing 29* (2020), 2766–2779. 1, 7, 8, 10

[LGG*17] LIN T.-Y., GOYAL P., GIRSHICK R., HE K., DOLLÁR P.: Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (2017), pp. 2980–2988. 2, 6

[LHC*20] LIU J.-J., HOU Q., CHENG M.-M., WANG C., FENG J.: Improving convolutional networks with self-calibrated convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), pp. 10096–10105. 3, 5

[LLZX21] LI H., LI J., ZHAO D., XU L.: Dehazeflow: Multi-scale conditional flow network for single image dehazing. In *Proceedings of the 29th ACM International Conference on Multimedia* (2021), pp. 2577–2585. 1

[LMB*14] LIN T.-Y., MAIRE M., BELONGIE S., HAYS J., PERONA P., RAMANAN D., DOLLÁR P., ZITNICK C. L.: Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2014), Springer, pp. 740–755. 3

[LMSC19] LIU X., MA Y., SHI Z., CHEN J.: Griddehazenet: Attention-based multi-scale network for image dehazing. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (2019). 1

[LOW*20] LIU L., OUYANG W., WANG X., FIEGUTH P., CHEN J., LIU X., PIETIKÄINEN M.: Deep learning for generic object detection: A survey. *International Journal of Computer Vision 128*, 2 (2020), 261–318. 2

[LPW*17] LI B., PENG X., WANG Z., XU J., FENG D.: Aod-net: All-in-one dehazing network. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (2017), pp. 4770–4778. 3, 7, 10

[LQS*19] LIANG X., QIU B., SU Z., GAO C., SHI X., WANG R.: Rain wiper: An incremental randomly wired network for single image deraining. In *Computer Graphics Forum* (2019), vol. 38, Wiley Online Library, pp. 159–169. 3

[LRF*18] LI B., REN W., FU D., TAO D., FENG D., ZENG W., WANG Z.: Benchmarking single-image dehazing and beyond. *IEEE Transactions on Image Processing 28*, 1 (2018), 492–505. 2, 6, 7

[LRY*22] LIU W., REN G., YU R., GUO S., ZHU J., ZHANG L.: Image-adaptive yolo for object detection in adverse weather conditions. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)* (2022), vol. 36, pp. 1792–1800. 1, 2, 3, 7, 8, 10

[PCS*19] PANG J., CHEN K., SHI J., FENG H., OUYANG W., LIN D.: Libra r-cnn: Towards balanced learning for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 821–830. 2

[QWB*20] QIN X., WANG Z., BAI Y., XIE X., JIA H.: Ffa-net: Feature fusion attention network for single image dehazing. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)* (2020), vol. 34, pp. 11908–11915. 3, 7, 10

[RCS*19] ROYCHOWDHURY A., CHAKRABARTY P., SINGH A., JIN S., JIANG H., CAO L., LEARNED-MILLER E.: Automatic adaptation of object detectors to new domains using self-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 780–790. 3

[RDGF16] REDMON J., DIVVALA S., GIRSHICK R., FARHADI A.: You only look once: Unified, real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 779–788. 2, 3

[RF17] REDMON J., FARHADI A.: Yolo9000: better, faster, stronger. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), pp. 7263–7271. 2, 3

[RF18] REDMON J., FARHADI A.: Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767* (2018). 2, 3

[RHGS15] REN S., HE K., GIRSHICK R., SUN J.: Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems 28* (2015). 2

[RLHS20a] REN D., LI J., HAN M., SHU M.: Not all areas are equal: A novel separation-restoration-fusion network for image raindrop removal. In *Computer Graphics Forum* (2020), vol. 39, Wiley Online Library, pp. 495–505. 3

[RLHS20b] REN D., LI J., HAN M., SHU M.: Scga-net: Skip connections global attention network for image restoration. In *Computer Graphics Forum* (2020), vol. 39, Wiley Online Library, pp. 507–518. 1

[RSA*21] REZAEIANARAN F., SHETTY R., ALJUNDI R., REINO D. O., ZHANG S., SCHIELE B.: Seeking similarities over differences: Similarity-based domain alignment for adaptive object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (2021), pp. 9204–9213. 2

[SCN21] SUN T., CHEN J., NG F.: Multi-target domain adaptation via unsupervised domain classification for weather invariant object detection. *arXiv preprint arXiv:2103.13970* (2021). 1

[SDVG18] SAKARIDIS C., DAI D., VAN GOOL L.: Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision 126*, 9 (2018), 973–992. 2, 6, 7

[SLC19] SHAN Y., LU W. F., CHEW C. M.: Pixel and feature level based domain adaptation for object detection in autonomous driving. *Neurocomputing 367* (2019), 31–38. 3

[SOYP20] SINDAGI V. A., OZA P., YASARLA R., PATEL V. M.: Prior-based domain adaptive object detection for hazy and rainy conditions. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2020), Springer, pp. 763–780. 1, 3

[SUHS19] SAITO K., USHIKU Y., HARADA T., SAENKO K.: Strong-weak distribution alignment for adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 6956–6965. 2

[SZB*22] SUN Z., ZHANG Y., BAO F., WANG P., YAO X., ZHANG C.: Sadnet: Semi-supervised single image dehazing method based on an attention mechanism. *ACM Transactions on Multimedia Computing, Communications, and Applications 18*, 2 (2022), 1–23. 3

[UVDSGS13] UIJLINGS J. R., VAN DE SANDE K. E., GEVERS T., SMEULDERS A. W.: Selective search for object recognition. *International Journal of Computer Vision 104*, 2 (2013), 154–171. 2

[WLW*20] WANG C.-Y., LIAO H.-Y. M., WU Y.-H., CHEN P.-Y., HSIEH J.-W., YEH I.-H.: Cspnet: A new backbone that can enhance learning capability of cnn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2020), pp. 390–391. 3, 4

[WYB*22] WANG Y., YAN X., BAO H., CHEN Y., GONG L., WEI M., LI J.: Detecting occluded and dense trees in urban terrestrial views with a high-quality tree detection dataset. *IEEE Transactions on Geoscience and Remote Sensing 60* (2022), 1–12. 1

[WYG*22] WANG Y., YAN X., GUAN D., WEI M., CHEN Y., ZHANG X.-P., LI J.: Cycle-snspgan: Towards real-world image dehazing via cycle spectral normalized soft likelihood estimation patch gan. *IEEE Transactions on Intelligent Transportation Systems* (2022), 1–15. 3

[YSP20] YASARLA R., SINDAGI V. A., PATEL V. M.: Syn2real transfer learning for image deraining using gaussian processes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), pp. 2726–2736. 9

[ZCM*20] ZHANG H., CHANG H., MA B., WANG N., CHEN X.: Dynamic r-cnn: Towards high quality object detection via dynamic training. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2020), Springer, pp. 260–275. 2

[ZPY*19] ZHU X., PANG J., YANG C., SHI J., LIN D.: Adapting object detectors via selective cross-domain alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 687–696. 3

[ZSL*20] ZHU X., SU W., LU L., LI B., WANG X., DAI J.: Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159* (2020). 5

[ZTHJ21] ZHANG S., TUO H., HU J., JING Z.: Domain adaptive yolo for one-stage cross-domain detection. In *Asian Conference on Machine Learning* (2021), PMLR, pp. 785–797. 3

[ZWK19] ZHOU X., WANG D., KRÄHENBÜHL P.: Objects as points. *arXiv preprint arXiv:1904.07850* (2019). 2

[ZZXW19] ZHAO Z.-Q., ZHENG P., XU S.-T., WU X.: Object detection with deep learning: A review. *IEEE transactions on Neural Networks and Learning Systems 30*, 11 (2019), 3212–3232. 2