



Article

Multi-Task Learning for UAV Aerial Object Detection in Foggy Weather Condition

Wenxuan Fang ^{1,†} , Guoqing Zhang ^{1,2,†} , Yuhui Zheng ¹ and Yuwen Chen ^{3,*}

¹ College of Computer Science, Nanjing University of Information Science and Technology, Nanjing 210044, China; fwenxuancontact@163.com (W.F.); guoqingzhang@nuist.edu.cn (G.Z.); zheng_yuhui@nuist.edu.cn (Y.Z.)

² College of Mathematical and Computational Sciences, Massey University, Auckland 0632, New Zealand

³ Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences, Chongqing 400714, China

* Correspondence: chenyuwen@cigit.ac.cn

† These authors contributed equally to this work.

Abstract: Adverse weather conditions such as haze and snowfall can degrade the quality of captured images and affect performance of drone detection. Therefore, it is challenging to locate and identify targets in adverse weather scenarios. In this paper, a novel model called Object Detection in a Foggy Condition with YOLO (ODFC-YOLO) is proposed, which performs image dehazing and object detection jointly by multi-task learning approach. Our model consists of a detection subnet and a dehazing subnet, which can be trained end-to-end to optimize both tasks. Specifically, we propose a Cross-Stage Partial Fusion Decoder (CSP-Decoder) in the dehazing subnet to recover clean features of encoder from complex weather conditions, thereby reducing the feature discrepancy between hazy and clean images, thus enhancing the feature consistency between different tasks. Additionally, to increase the feature modeling and representation capabilities of our network, we also propose an efficient Global Context Enhanced Extraction (GCEE) module to extract beneficial information from blurred images by constructing global feature context long-range dependencies. Furthermore, we propose a Correlation-Aware Aggregated Loss (CAALoss) to average noise patterns and tune gradient magnitudes across different tasks, accordingly implicitly enhancing data diversity and alleviating representation bias. Finally, we verify the advantages of our proposed model on both synthetic and real-world foggy datasets, and our ODFC-YOLO achieves the highest mAP on all datasets while achieving 36 FPS real-time detection speed.

Keywords: UAV images; object detection; YOLO; foggy weather condition



Citation: Fang, W.; Zhang, G.; Zheng, Y.; Chen, Y. Multi-Task Learning for UAV Aerial Object Detection in Foggy Weather Condition. *Remote Sens.* **2023**, *15*, 4617. <https://doi.org/10.3390/rs15184617>

Academic Editors: Lorenzo Capineri and Chiman Kwan

Received: 30 August 2023

Revised: 10 September 2023

Accepted: 15 September 2023

Published: 20 September 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Object detection is a fundamental problem in computer vision and has numerous practical applications, such as in autonomous driving [1], Pedestrian re-identification [2–5] and UAV Aviation Inspection [6]. For example, object detection [7–9] can judge each identifiable object and locate the position on the images, which assists an autonomous driving perception system to navigate safely in complicated driving environments. However, due to the influence of complex weather conditions, the degradation of image quality can negatively impact feature extraction and analysis in object detection tasks. This also causes detectors trained on clean datasets to fail on these degraded images, posing a threat to the safety of autonomous driving. Therefore, improving the generalization performance of detectors under complex weather conditions has become an attractive research problem.

In recent years, object detection methods [10–12] based on Convolutional Neural Networks (CNNs) have achieved remarkable performance. However, these methods only consider simple scenarios, when the environment is relatively harsh (e.g., fog and

rain), their performance perform suboptimally. Figure 1a exhibits the detection results of YOLOx [12], it can be seen that the missed detections are obvious in dense fog.

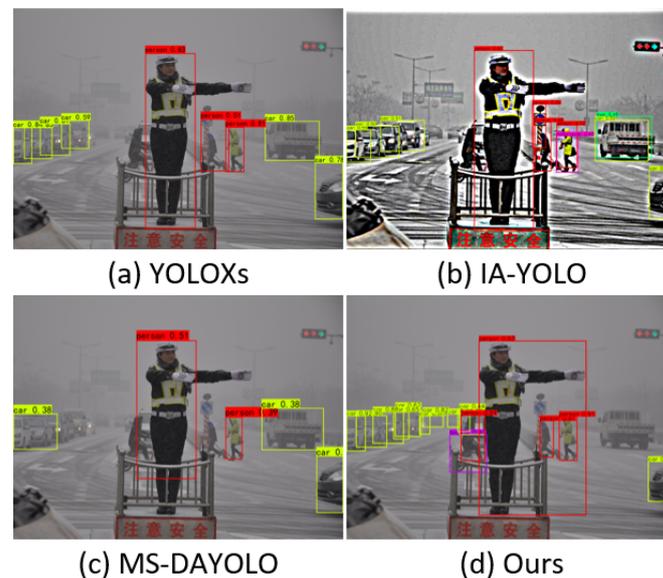


Figure 1. Comparing of the detection results of different methods in real-world foggy environments: (a) YOLOx, (b) IA-YOLO, (c) MS-DAYOLO, (d) the proposed method. The objects of interest, including people, cars, motorbikes, buses, and bicycles, are highlighted in red, green, blue, dark green, and purple, respectively (best viewed in color).

Recent, existing methods such as image dehazing [13,14] and image adaptation [15] are used as a pre-processing step in detection methods to improve degraded image quality or remove impurities in an image. However, improvements in image quality does not always improve the performance of detection network [14]. As shown in Figure 1b, it can be clearly observed that although the image adaptation detection method IA-YOLO [16] suppresses the impact of specific weather condition on an image, but it also suffers from useful information loss and causes false and missed detection in foggy weather, such as the bicycle is falsely detected as a pedestrian (see purple bounding box) and the cars on the left in this image are not fully detected.

Some researchers [17–19] regard the problem of object detection in a foggy weather as a domain adaptation task, where the goal is to align the features of clean images (source domain) with those of degraded images (target domain) using domain adaptation techniques. Since domain-adaptive methods usually focus on the distribution of data across different domains rather than changes in object appearance and corresponding information loss, these methods may not be able to fully exploit additional information in degraded images that is beneficial for detection. As shown in Figure 1c, it is evident that MS-DAYOLO [19] can only detect objects with salient features, while objects heavily occluded by fog cannot be detected. In contrast, our approach can detect more objects with higher scores in foggy images, as shown in Figure 1d.

To tackle the aforementioned challenges, we propose a multi-task Object Detection method in a Foggy Condition with YOLO (ODFC-YOLO), which can simultaneously perform object detection and image dehazing in a unified end-to-end trainable framework. Our framework is composed of two subnets: a detection subnet and a dehazing subnet, and the architecture is shown in Figure 2. Specifically, we first propose a Cross-Stage Partial Fusion Decoder (CSP-Decoder) to effectively removes weather-related information from the shared features, thereby strengthen the feature consistency across tasks. In addition, to improve the feature modeling ability of our model, we also propose a Global Context Enhancement Extraction (GCEE) module, which can establish the global context feature long-range dependencies to extract extra beneficial information for detection from degraded

images. Finally, we propose a Correlation-Aware Aggregated Loss (CAALoss) to train our network in a way that averages noise patterns and tunes gradient magnitudes of different tasks that implicitly enhancing data diversity and alleviating the representation bias. Extensive experiments on several representative datasets show that our ODFC-YOLO method obtains the best detection accuracy and maintains real-time performance compared to other representative advanced detection methods.

In brief, our work makes the following contributions:

- We propose a multi-task learning architecture that simultaneously performs image dehazing and object detection in foggy weather scenarios and allows for end-to-end training.
- We propose a Cross-Stage Partial Fusion Decoder (CSP-Decoder) to remove weather-related information and produce clean features, which can reduce the difference between hazy and clean images.
- We propose a Global Context Enhancement Extraction (GCEE) module, which extracts additional information from degraded images through the construction of global context feature long-range dependencies, improving the robustness of the network as well as the ability to handle complex weather conditions.
- We design a novel Correlation-Aware Aggregated Loss (CAALoss) to average noise patterns and tune gradient magnitudes between different tasks to implicitly strengthen data diversity and mitigates representation bias.
- We conduct extensive experiments to verify the superiority of our proposed ODFC-YOLO over several state-of-the-art methods. Also, we provide a detailed analysis and discussion to comprehensively study the robustness of our method in harsh weather conditions.

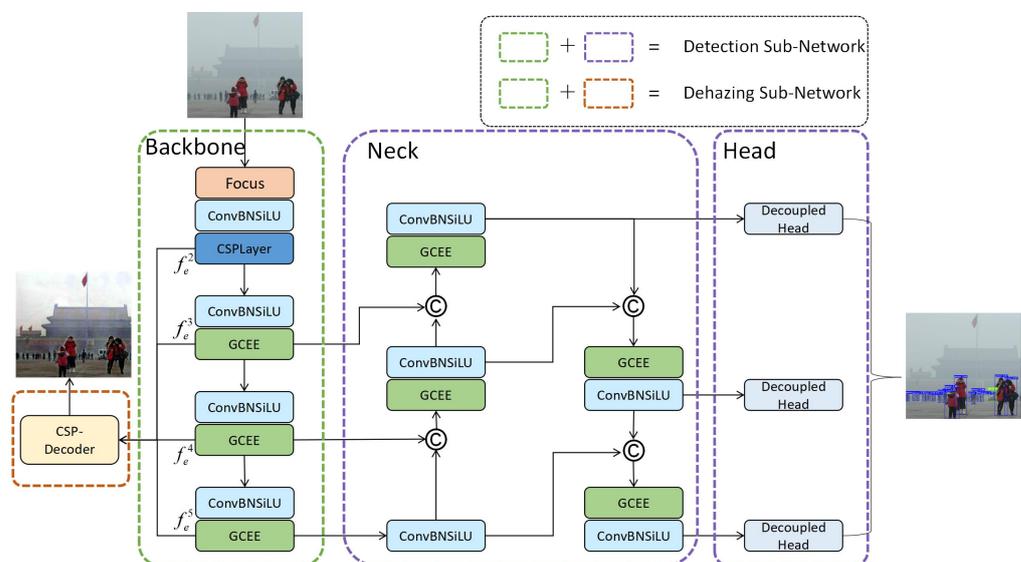


Figure 2. The architecture of our ODFC-YOLO, which is an end-to-end multi-task learning-based framework, which mainly includes a dehazing subnet and a detection subnet. It's worth to mention that the dehazing subnet only participates during training, which generates the clean images, but these images are not used as input to the detection subnet in the inference time. The green dotted line plus the orange dotted line indicates the dehazing subnet; the green dotted line plus the purple dotted line indicates the detection subnet. CSP-Decoder: Cross-Stage Partial Fusion Decoder.

2. Related Works

2.1. Object Detection

Object detection is a well-studied vision task that aims to detect and locate objects of interest in an image and classify them. As a popular subject, it can be divided into two main types [20]: two-stage methods and one-stage methods.

Two-stage methods: Two-stage methods [21–26] are designed as a coarse-to-fine process that involves generating a large set of low-quality proposals, which are then refined through further feature extraction and prediction. Early two-stage methods mainly used handcrafted features, which made them difficult to develop and improve. With the advancement of deep learning technology, Girshick et al. [21] introduced R-CNN and employed Selective Search (SS) [27] to generate a number of candidate boxes called “proposals” and then extracted features for each proposal using CNNs, followed by SVM classification. Afterwards, Faster-RCNN [23] proposed a fully convolutional Region Proposal Network (RPN) is utilized to generate fewer but higher-quality proposals, reducing detector complexity while improving performance and speed. Although the region proposal mechanisms boosts the performance of two-stage detectors, it also limits their inference speed.

One-stage methods: On the other hand, one-stage methods [10,12,28–32] aim to predict the detection results directly from the whole image without generating proposals, which greatly simplify the pipeline of the detector, and achieve much faster inference speed than two-stage methods. YOLO [29] was one of the first one-stage detectors, which formulated object detection as a regression problem by directly predicting the classification confidence and bounding box offsets for each object in the image. Subsequent series of improvements based on YOLO have significantly improved detection speed and accuracy [10,12,28,30,31].

Apart from this, some recent works have opened new avenues for object detection [33–39]. For example, Ref. [35] exploited the frequency characteristics of transmitted signals for radio frequency (RF)-based drone detection and classification. We propose a novel machine learning (ML) framework for detecting and classifying ADr sounds from a variety of sounds such as birds, airplanes, and thunderstorms in noisy environments. Anwar et al. [36] proposed a novel machine learning (ML) framework for detecting and classifying ADr sounds from a variety of sounds such as birds, airplanes, and thunderstorms in noisy environments. Huang et al. [37] proposed a new method for object detection in UAV images that the multi-agent detection network with unified foreground encapsulation (UFPMP-Det), which clusters the sub-regions given by the coarse detector and suppresses the background. Furthermore, Carion et al. [39] first proposed an end-to end object detector based on Vision Transformer, called DETR (DEtection TRansformer), which simplifies the detection pipeline and directly predicts one-to-one object sets using bipartite matching. Subsequently, many DETR variants were proposed to improve the performance of detectors. For example, Deformable-DETR [38] accelerates the convergence of training with multi-scale features by improving the efficiency of the attention mechanism. Anchor-DETR [33] reduces the difficulty of query optimization. Lite-DETR [34] improves detection performance by updating high-level features and low-level features in an interleaved manner.

2.2. Image Dehazing

Image dehazing is a critical task in computer vision that aims to remove haze or fog from images to improve their visual clarity and quality. Recently, significant efforts have been made in developing effective dehazing methods. For example, Ye et al. [40] designed a novel Separable Hybrid Attention (SHA) module and density map to effectively capture the degradation caused by non-uniform distributions at the feature level. Guo et al. [41] introduced the haze density prior as an absolute position embedding into the transformer. Bai et al. [42] proposed a deep pre-dehazer to generate intermediate consequences as reference images and explores the guidance information within these images. Song et al. [43] modified the key structure of the Swin transformer to adapt image dehazing. Lu et al. [44] proposed an enhanced parallel attention module to efficiently handle an uneven haze distribution.

2.3. Object Detection in Adverse Weather

Object detection in adverse weather conditions is a challenging task. One intuitive method is to pre-process the degraded images using existing restoration methods, such as image enhancement, dehazing and blurring methods, to remove weather-specific information. IDOD-YOLOV7 [45] proposed an image defogging module (IDOD), which performs image enhancement and combined with YOLOV7 in a weakly supervised manner. An alternative approach is to use multi-task learning to simultaneously address image restoration and detection tasks, which can help alleviate the effect of adverse weather information on the detection performance. Li et al. [46] proposed a joint architecture network integrating the deblurring and detection modules for end-to-end processing, and designed an interval iterative data refinement training strategy to facilitate the learning of the deblurring module without supervision. Recent methods [17–19] attempt to improve detector performance from the perspective of domain adaptation. Furthermore, some studies [47–49] collected more datasets with complex weather conditions to improve the robustness of detectors under such scenarios.

2.4. Multi-Task Learning

Multi-task learning methods [50–52] aimed to learn multiple related tasks simultaneously, benefiting other tasks through the knowledge gained from one task. One popular method is the Task Relation Network [52] (TRN), which built explicit relationship between tasks from a statistical perspective. However, TRN regards all tasks as equally related and when a certain task has stronger correlations, it will cause gradients to a point in the wrong direction. In this work, differing from the methods proposed in [52], our method establishes implicit task relationships between different tasks to maintain feature consistency for better joint optimization.

3. Methodology

3.1. Overview of the Proposed ODFC-YOLO

The main objective is to jointly optimize image dehazing and object detection tasks by multi-task learning. ODFC-YOLO consists of two independent subnets that can seamlessly exchange information and learn from each other during training. On the one hand, the two subnets share the feature extraction network, and the proposed CSP-Decoder enhances the consistency of shared features. On the other hand, the proposed novel Correlation-Aware Aggregated Loss (CAALoss) can average noise patterns and tune gradient magnitudes between different tasks to better capture the implicit correlations between tasks, thereby implicitly enhancing data diversity and mitigating representation biases. Finally, we chose the YOLOX-small (YOLOXs) detector as the basis of our detection subnet. In order to improve the feature extraction capability of the detection network in severe weather conditions, we also propose the GCEE module to obtain contextual long dependencies to facilitate this purpose.

The overall structural pipeline of ODFC-YOLO is shown in Figure 2. Firstly, given a foggy image as input, we split the image into equal-sized patches by a focus operation, and then recombine the patches to strengthen image features. Then, a cross-stage partial module [30] and multiple global context enhancement extraction modules are employed to extract salient object representations, which are transferred to the Cross-Stage Partial Fusion Decoder (CSP-Decoder) and neck module, which perform corresponding tasks respectively. In this way, CSP-Decoder produces clean features and balances the roles of subnets by optimization strategy, while the neck module combines shallow and deep features to enrich the features. Finally, the detection head module benefits from the joint learning framework to produce the more accurate scores, categories and locations of the detected object. Remarkably, the dehazing subnet is only involved during training and turned off during inference.

3.2. Dehazing Sub-Network

In our proposed ODFC-YOLO, the dehazing subnet plays a crucial role in removing haze from images, which is achieved by an encoder-decoder structure as depicted in Figure 2. The detection subnet's backbone network serves as the encoder to extract latent features from the hazy image, while the proposed CSP-Decoder is utilized to remove haze and restore clean image.

Due to the different update directions of these two tasks, the features extracted by the backbone network may not be optimal for both tasks. To address this issue, we propose CSP-Decoder that aims to decrease the discrepancies between the features extracted by the backbone network for both tasks and improve the image dehazing performance. Additionally, the dehazing subnet improves the range of the receptive field, which helps to deliver the training data from low-quality domain to high-quality domain, thus effectively suppressing the domain shifts (i.e., the inconsistency between the training domain and the actual environment) that exists in the training images.

Our dehazing subnet is composed of an encoder, and our proposed Cross-Stage Partial Fusion Decoder (CSP-Decoder). The encoder aims to extract latent features of the haze image and our CSP-Decoder is responsible for reducing the effects of weather-specific information on the extracted features, thus reducing the domain shift between the training and application environments.

As shown in Figure 3, the CSP-Decoder contains five cross-stage partial layers, three SKFusion layers, and four upsampling operations. The input image is reduced by $1/32$ after passing through the backbone network, and the pixelShuffle layer is employed to increase the feature resolution such that the output image resolution is kept the same as the input image. The CSP-Decoder accepts outputs from four stages of the encoder (denoted as $f_e^i, i \in [2, 3, 4, 5]$). Among them, the shallow layers of f_e^2, f_e^3 , and f_e^4 from encoder contain features of different resolutions respectively, which are beneficial for enhancing deeper visibility features. The SKFusion layer [43] is used to fuse multiple branch features and alleviate the gradient disappearance to facilitate the dehazing task. Finally, the clean image is restored by a 1×1 convolution operation. We attempt to replace the cross-stage partial layers with GCEE module in CSP-Decoder, but it had no significant effect on detector performance.

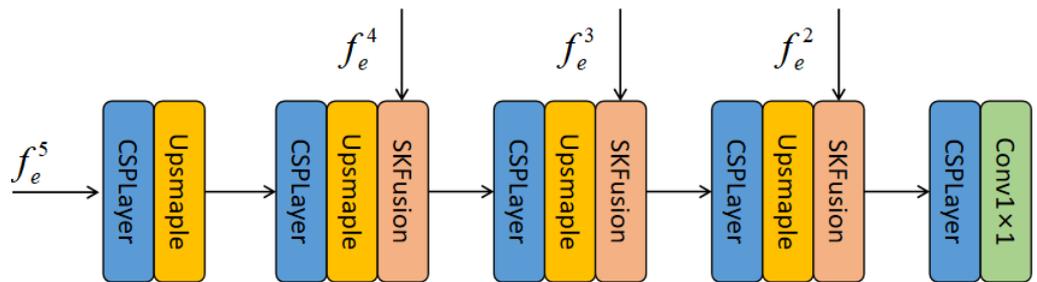


Figure 3. The architecture of Cross-Stage Partial Fusion Decoder (CSP-Decoder), which removes the negative impact of degraded images on feature extraction and produces high-quality clean features.

It is worth noting that the primary objective of the dehazing subnet is not to produce clear images for the input of the detection subnet, but rather to enhance the feature extraction capability of the backbone network through multi-task optimization.

To train the dehazing subnet, we apply L1 loss to the network:

$$\mathcal{L}_{re} = \frac{1}{N} \sum_{i=1}^N \|\hat{y}_i(x) - y_i(x)\|, \quad (1)$$

where N refers to the number of training samples, $\hat{y}(x)$ and $y(x)$ express the estimated clean image and corresponding ground-truth image, respectively. Actually, considering

that complex loss functions will bring unnecessary training burden, we prefer to choose simple L1 loss for better performance balance.

3.3. Detection Sub-Network

We choose YOLOX as our detection subnet, which boasts a lightweight architecture while maintaining strong performance on various benchmark datasets. However, like many other existing object detectors, YOLOX also suffers from performance degradation when it encountering harsh weather conditions.

Our detection subnet is composed of three key components: backbone network, neck module, and head module. The backbone network is based on the feature extraction architecture of YOLOV4 [30] and we replace the cross-stage partial module with our proposed GCEE module to obtain better feature modeling capabilities. To augment the localization capability of the feature representation, the neck module employs top-down feature fusion and horizontal connections strategy, which improves the accuracy of detecting small objects in hazy scenes. Lastly, the head module is designed to independently perform classification and localization of objects at multiple scales.

Traditional Convolutional Neural Networks use deep module stacks to enhance the receptive fields and extract contextual high-level semantic information. However, this design is computationally expensive and may produce redundant features that do not contribute to the final output. To this end, we propose the Global Context Enhanced Extraction (GCEE) module, which is designed to expand the regional receptive field and capture long-range contextual dependencies while improving computationally efficient. The GCEE module enhance the representation capabilities of the detection subnet by selectively extracting the most useful features, while reducing computation cost and avoiding unnecessary feature generation.

In Figure 4, our GCEE module first uses two local awareness branches to increase the influence regions, which improves the model's ability to transform features. Next, we gather information from different local perspectives to create global awareness features. To establish long-range dependencies, GCEE module further enhances the original features using an attention-like mechanism. Specifically, the context of the globally aware features is calculated by the global interactions of all pixel values using pixel-level dot products of linear units (1×1 convolutions and *softmax* layers). Furthermore, only two linear projections and a LayerNorm layer are utilized to improve feature diversity. Finally, the global awareness features are reweighted to the position of each context feature to obtain complex and long-term global context dependencies. In this way, by associating the input features with attention features that have long dependencies to perceive local objects, our network can extract more information that is beneficial to detection due to its better feature extraction ability, as demonstrated in Figure 5.

To train the detection subnet, we employ the loss function consisting of three parts, which can be expressed as:

$$\mathcal{L}_{de} = \frac{\lambda L_{reg} + L_{cls} + L_{obj}}{N_{pos}}, \quad (2)$$

where L_{reg} is the bounding box regression loss, which is calculated using IOU loss [53]. The coefficient factor λ is used to adjust the proportion of L_{reg} , and it is set to 5.0 in our experiment. L_{cls} and L_{obj} are the classification loss and confidence loss, respectively, which are calculated using binary cross-entropy loss (BCELoss). N_{pos} refers to the number of positive samples. Notably, only positive samples are used for calculation in L_{cls} and L_{reg} .

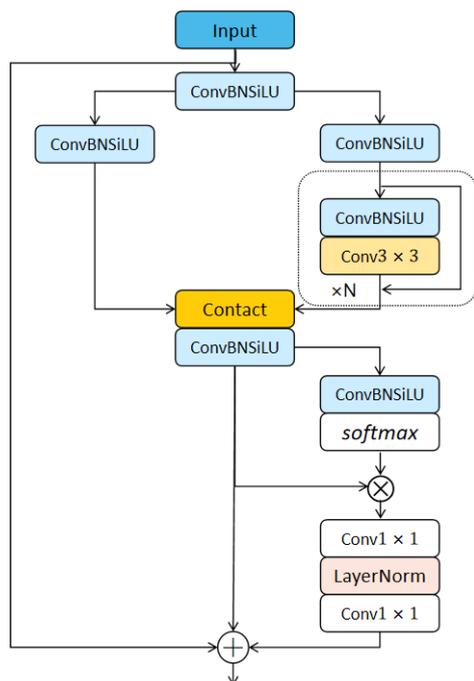


Figure 4. The Global Context Enhanced Extraction (GCEE) module, which is a mechanism for improving the feature extraction capability of our model by building global feature long-range dependencies.

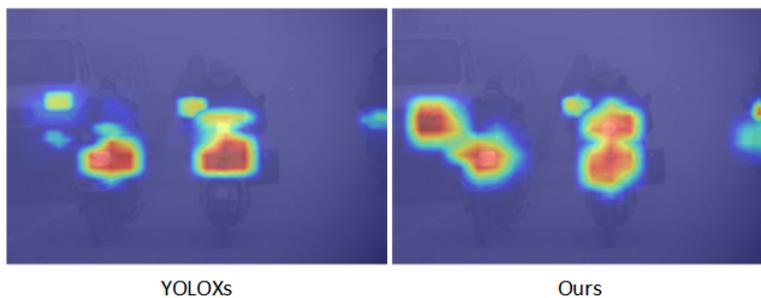


Figure 5. Visualization of regions of attention for our Global Context Enhanced Extraction (GCEE) module.

3.4. Correlation-Aware Aggregated Loss

Due to the different data distributions and noise patterns of multiple tasks, we weight the relative contribution of each task such that they have equal importance, and add abnormal noise to improve the generalization ability of our model. The CAALoss is calculated as:

$$\mathcal{L}_{caa} = \lambda_1 \mathcal{L}_{de} + \lambda_2 \mathcal{L}_{re} + \epsilon, \tag{3}$$

where \mathcal{L}_{de} is the object detection loss and \mathcal{L}_{re} refers to the dehazing loss. The weighting factors λ_1 and λ_2 are utilized to balance the importance of these two terms. To further improve the model’s ability to handle outliers, a small amount of noise is added to the loss function, which is denoted as ϵ . This noise term is calculated as $\epsilon = 0.02 * \lambda_1 + \lambda_2$, so that the model is not too sensitive to outliers. Using a unified loss function with different weighting factors allows the network to focus on reconstructing clean features, while taking into account the importance of the detection task. The best performance was observed experimentally for $\lambda_1 = 0.2$ and $\lambda_2 = 0.8$.

4. Experiments

To estimate the performance of our ODFC-YOLO with other advanced detection methods in severe weather conditions, we conduct extensive qualitative and quantitative experiments and our network is trained using VOC-Foggy dataset. The

Foggy Driving dataset [49], VisDrone [6] UAV aerial photography dataset and Real-world Task-driven Testing Set (RTTS) [14] are used for testing sets. All experiments are conducted using the Pytorch framework and run on 2 NVIDIA GeForce RTX 3090 GPUs for better performance.

4.1. Datasets

Considering the limited number of publicly datasets available for object detection under severe weather conditions, inspired by the Liu et al. [16], we generate a synthetic foggy dataset, called VOC-Foggy by adding haze to the PASCAL-VOC dataset. The atmospheric scattering model is used to generate the foggy images, and it is mathematically represented as follows:

$$I(x) = J(x)e^{-\beta d(x)} + A(1 - e^{-\beta d(x)}), \quad (4)$$

where $I(x)$ refers to the hazy images and $J(x)$ is the clean images, the global atmospheric light component, denoted by A , which is set to 0.5. $e^{-\beta d(x)}$ represents the atmospheric transmission coefficient, β is set to $0.01 * i + 0.05$, and i can take an integer between 0 and 9. The value of $d(x)$ is given by:

$$d(x) = -0.04 * \rho + \sqrt{\max(\text{row}, \text{col})}, \quad (5)$$

where ρ denotes the Euclidean distance between the current position and the center pixel, and row and col correspond to the number of rows and columns in the image, respectively.

Table 1 presents the total number of images and instances per class for the used dataset. The RTTS dataset contains annotated real-world haze images 5 categories (cars, people, buses, bicycles, motorcycles), with a total of 4322 images. It is currently the largest real-world foggy object detection dataset. The Foggy Driving dataset consists 101 annotated images that depicting real-world foggy driving scenarios with 8 categories (car, bus, train, truck, bicycle, person, rider and motorcycle). In order to maintain the consistency of our experiments, we only select the same 5 categories for testing in the Foggy Driving dataset as in the RTTS dataset.

Table 1. For statistics on the total number of images and the number of instances per class for all datasets used, including voc-fog-tv (VOC-Foggy-train), voc-fog-ts (VOC-Foggy-test), RTTS and FoggyD (Foggy Driving dataset). We select only five classes from the dataset: car, person, bus, bicycle (bic), motorcycle (motc).

Dataset	Images	Person	Bic	Car	Bus	Motc	Total
voc-fog-tv	8111	13,256	1064	3267	822	1052	19,561
voc-fog-ts	2734	4528	377	1201	211	325	6604
RTTS	4322	7950	534	18,413	1838	862	29,577
FoggyD	101	269	17	425	17	9	737

4.2. Training Settings

We employ the SGD optimizer with an initial learning rate of 1×10^{-2} to train our proposed ODFC-YOLO model. The total training epochs and batch size are to 100 and 32, respectively and the learning rate is adjusted dynamically during training using a cosine annealing decay strategy. Additionally, we do not use image augmentation techniques, commonly used in YOLO-based models, since these techniques would increase the difficulty of reconstructing images for the dehazing subnet, resulting in poor performance of the object detector. We set the image size for training and testing to 640×640 pixels.

4.3. Compare with SOTA Methods

We compare our method with several advanced methods, including: (1) “dehaze + detect” pre-processing methods, which use a dehazing algorithm as a pre-processing step, followed by a detector such as YOLOX. We choose four representative dehazing methods, including DCP [54], AOD-Net [55], GCA-Net [56], and FFA-Net [57], all of which are trained on the ITS dataset [14]; (2) DS-Net [50] based on multi-task learning; (3) MS-DAYOLO [19] based on domain adaptation; (4) IA-YOLO [17] based on image adaptation. We retrained all the compared methods (except for the pre-processing methods) on the VOC-Foggy-train dataset. It is noteworthy that we trained YOLOX and our ODFC-YOLO from scratch without using any pre-trained weights.

As described in Table 2, our proposed ODFC-YOLO outperforms other compared methods and obtains the highest mAP on the VOC-Foggy-test dataset, especially brings 8% improvement over YOLOXs. This indicates that our proposed method can improve the detection performance in severe weather conditions by jointly learning. Specifically, our method outperforms all compared methods in three out of five categories except the results obtained in a person and bus categories. The main reason is that in order to reduce the training complexity of the dehazing subnet, we did not employ data augmentation technology, thus sacrificing the diversity of training samples and affecting detection subnet performance. We also see that ODFC-YOLO surpasses all multi-task learning approaches. Although TogetherNet [51] obtains competitive results in two out of five categories, our method still outperforms it and achieves a 2% improvement in mAP. This proves that enhancing the feature consistency of different tasks is more conducive to the improvement of detection performance.

Table 2. Performance comparisons of our ODFC-YOLO with other advanced detection methods on the RTTS dataset. The mAP (mean Average Precision) used to evaluate object detection performance.

Methods	Person	Bicycle	Car	Motorbike	Bus	mAP
YOLOXs(arXiv’21) [12]	80.81	74.14	83.63	75.35	86.40	80.07
YOLOXs*(arXiv’21) [12]	79.97	67.95	74.75	58.62	83.12	72.88
AOD-YOLOXs*(ICCV’17) [55]	81.26	73.56	76.98	71.18	83.08	77.21
DCP-YOLOXs*(TPAMI’10) [54]	81.58	78.80	79.75	78.51	85.64	80.86
GCA-YOLOXs*(WACV’19) [56]	81.50	<u>80.89</u>	84.18	78.42	77.69	80.53
FFA-YOLOXs*(AAAI’20) [57]	78.30	70.31	69.97	68.80	80.72	73.62
MS-DAYOLO(ICIP’21) [19]	82.52	75.62	<u>86.93</u>	81.92	<u>90.10</u>	83.42
DS-Net(TPAMI’21) [50]	72.44	60.47	81.27	53.85	61.43	65.89
IA-YOLO(AAAI’22) [16]	75.14	67.84	76.91	57.91	67.61	72.03
TogetherNet(CGF’22) [51]	87.62	78.19	85.92	<u>84.03</u>	93.75	<u>85.90</u>
BAD-Net(TPAMI’23) [46]	-	-	-	-	-	85.58
ODFC-YOLO(Ours)	<u>86.67</u>	88.83	91.16	86.60	86.85	88.02

In order to assess the robustness of our proposed method on real-world scenarios, we evaluated it on two challenging datasets: the RTTS and Foggy Driving datasets. The evaluation of performance is based on the mean average precision (mAP) metric, and the outcomes of the compared methods are depicted in Tables 3 and 4. We can obtain a similar conclusion to the previous section that our method acquires the highest mAP on these two datasets.

To provide an intuitive understanding of the detection performance of various methods, we also visualize the detection results of our model and other representative methods on the Foggy Driving (Figure 6) and RTTS (Figure 7) datasets. We can observe that our method can detect a greater quantity of objects with higher accuracy even in challenging scenarios with dense fog, which is attributed to our designed architecture that considers both global context information and local spatial detail. It can also observe from Figure 7 that although the image-adaptive method IA-YOLO [16] can detect a certain number of targets, it is still struggles with identifying hidden objects in dense fog. In addition, our model also maintains excellent detection performance on low-light images without training

on professional dataset, as shown in Figure 8. It can be found that ODFC-YOLO can adapt to different lighting conditions, which is crucial for target detection in various scenarios.



Figure 6. Detection results of our ODFC-YOLO and other advanced detection methods on Foggy Driving dataset. Obviously, the proposed ODFC-YOLO can detect more objects with higher confidence.

Table 3. Performance comparisons of our ODFC-YOLO with other advanced detection methods on the RTTS dataset.

Methods	Person	Bicycle	Car	Motorbike	Bus	mAP
YOLOXs	<u>80.88</u>	64.27	56.39	52.74	29.60	56.74
YOLOXs*	76.28	60.44	64.73	50.06	24.86	55.04
AOD-YOLOXs*	77.26	62.43	56.70	53.45	30.01	55.83
GCA-YOLOXs*	79.12	67.10	56.41	<u>58.68</u>	34.17	58.64
DGP-YOLOXs*	78.69	<u>67.99</u>	55.50	57.57	33.27	58.32
FFA-YOLOXs*	77.12	66.51	64.23	40.64	23.71	52.64
MS-DAYOLO	74.22	44.13	70.91	38.64	<u>36.54</u>	57.39
DS-Net	68.81	18.02	46.13	15.15	15.44	32.71
IA-YOLO	67.25	35.84	42.65	22.52	17.64	37.89
TogetherNet	82.70	57.27	75.31	55.40	37.04	<u>61.55</u>
Ours	79.63	68.71	<u>74.26</u>	68.41	34.23	62.05

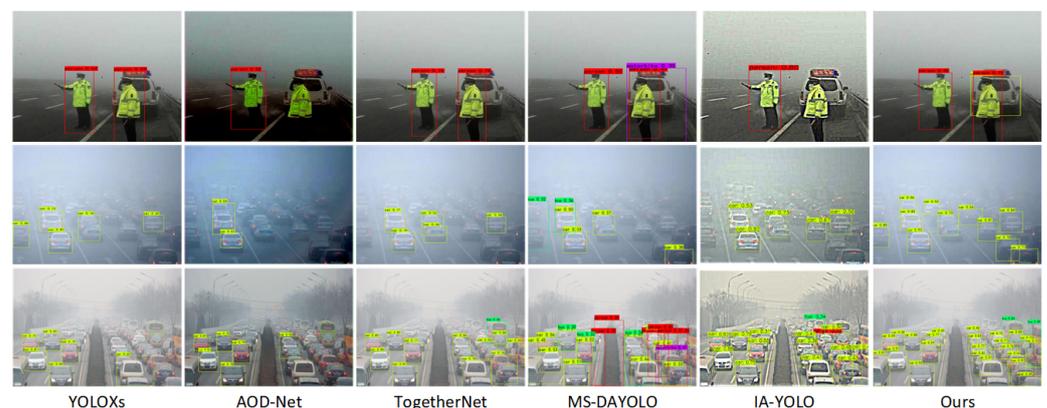


Figure 7. Detection results of our ODFC-YOLO and other advanced detection methods on real-world foggy dataset (RTTS).

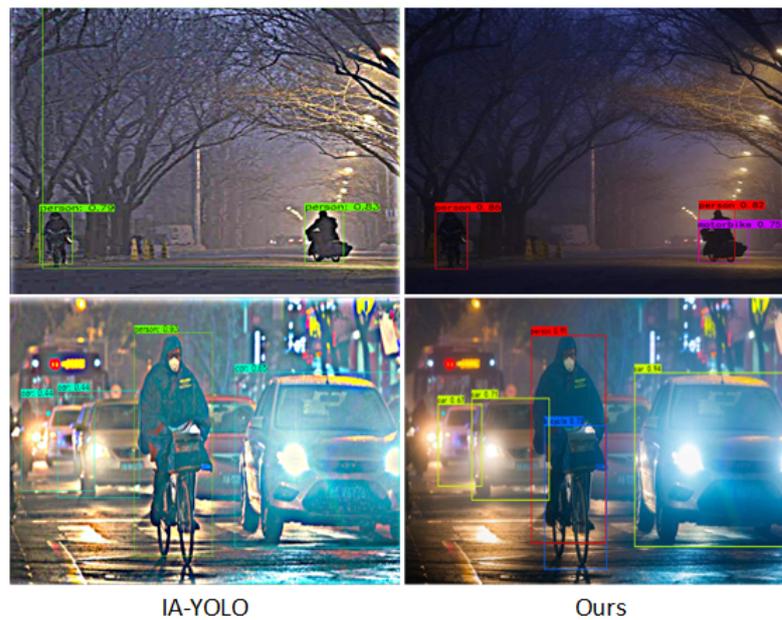


Figure 8. Detection results of our method and representative image adaptation methods (IA-YOLO) in low-light conditions.

Table 4. Performance comparisons of our ODFC-YOLO with other advanced detection methods on the Foggy Driving dataset.

Methods	Person	Bicycle	Car	Motorbike	Bus	mAP
YOLOXs	24.36	27.25	55.08	8.04	44.79	33.06
YOLOXs*	26.58	23.67	56.22	6.74	41.87	32.49
AOD-YOLOXs*	26.15	33.72	56.95	6.44	34.89	32.51
GCA-YOLOXs*	27.96	<u>34.11</u>	56.36	6.77	34.21	33.77
DCP-YOLOXs*	22.64	11.07	56.37	4.66	36.03	31.56
FFA-YOLOXs*	19.22	21.40	50.64	3.69	43.85	28.74
MS-DAYOLO	21.52	34.57	<u>57.41</u>	18.20	46.75	34.89
DS-Net	26.74	20.54	54.16	7.14	36.11	29.74
IA-YOLO	20.24	19.04	50.67	8.11	22.97	25.70
TogetherNet	<u>30.48</u>	30.47	57.87	14.87	40.88	<u>36.75</u>
Ours	35.69	35.26	59.15	<u>16.17</u>	<u>45.88</u>	38.41

4.4. Experimental Results on the VisDrone Dataset

We select the VisDrone [6] UAV aerial photography dataset for showcasing the advantages of our model. Employing the same equal.4 approach, we introduce haze to aerial images. The outcomes are prominently display in Figure 9, where the focus lies on detecting objects within fog-laden scenes, including both daytime and nighttime aerial images in foggy conditions. A discernible observation emerges from the results: the YOLOXs model encounters count of missed detections due to the occlusive nature of the fog. In stark contrast, our proposed model excels in object detection, even amid such obstructive atmospheric conditions. Notably, our model not only improves object detection rates but also effectively distinguishes foreground objects from background elements, showcasing its robust performance.



Figure 9. The detection results of the VisDrone dataset under different scenes.

4.5. Comparison on Rainy Images

To verify the generalization ability of our model in other severe weather conditions, we choose the RainCitySpaces [48] dataset to evaluate our model's performance. Compared to the Rainy WCity [47] dataset, RainCitySpaces is widely used for object detection task in rainy day and has more train images. We randomly choose 2500 images for training and 450 images for testing, and focus on detecting the same five object categories as in previous experiments.

The comparison results of our ODFC-YOLO with YOLOXs [12], GCA-YOLOX [56], IA-YOLO [16], MS-DAYOLO [19] and TogetherNet [51] on the RainCitySpaces dataset are presented in Table 5 and our method achieves the highest mAP. The second column of Table 5 lists the detection results without retraining on the dataset. Our method obtains 6.99%, 5.72%, 27.29%, 2.05% and 7.14% improvements over YOLOXs, GCA-YOLO, IA-YOLO, MS-DAYOLO and TogetherNet, respectively. In addition, the performance is significantly improved after retraining the model. Figure 10 shows the detection results of all compared methods, we can see that our method can detect more objects under the rainy and foggy weather conditions, indicating its stronger generalization potential without specialized training.

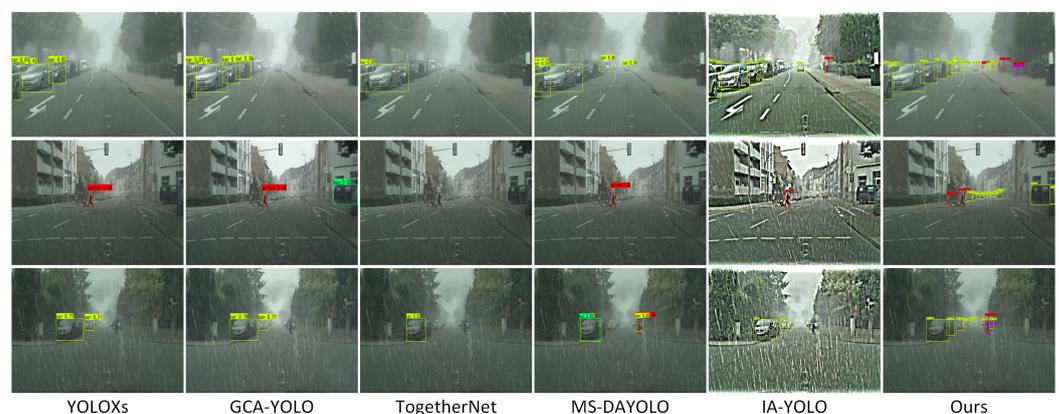


Figure 10. Detection results of our ODFC-YOLO and the other advanced detection methods without retrained on the RainCityscapes dataset.

Table 5. Performance comparison of our ODFC-YOLO with other detection methods on the RainCITYscapes dataset. The mAP (mean Average Precision) used to evaluate object detection performance.

Methods	Without-Retrained	retrained
YOLOXs [12]	35.49	40.62
GCA-YOLOXs [56]	36.76	39.15
IA-YOLO [16]	14.52	15.32
MS-DAYOLO [19]	40.43	44.37
TogetherNet [51]	35.34	39.14
Ours	42.48	48.20

5. Ablation Study

5.1. Effects of Different Components in ODFC-YOLO

To further investigate the contribution of individual components and its internal structure in our ODFC-YOLO model, Table 6 reports a set of ablation experiments on the RTTS dataset with the same experimental configuration. We consider different variants of the model such as CSP-Decoder* (without SKFusion layer), CSP-Decoder, and Global Context Enhancement Extraction (GCEE) module. By reassembling components in different ways, we are able to determine the optimal model performance.

Table 6. Ablation study of different combination strategies of the proposed modules are performed on the RTTS dataset. Decoder represents the version of CSP-Decoder with SKFusion module, and Decoder* without SKFusion module.

Modules	Base	Variants	Variants	Variants	Variants	Variants
Decoder*	w/o	✓	w/o	w/o	✓	w/o
Decoder	w/o	w/o	✓	w/o	w/o	✓
GCEE	w/o	w/o	w/o	✓	✓	✓
mAP	56.74	59.06	59.76	57.76	60.65	62.05

Given that the SKFusion module has demonstrated excellent performance in image restoration, we add it to our proposed ODFC-YOLO model to enhance the effect of the decoder. The experimental results demonstrate that the addition of the SKFusion module to CSP-Decoder significantly improves our model's performance. It can be seen that each component of our ODFC-YOLO contributes to the improvement of the detection performance. The proposed CSP-Decoder boosts the performance by 2% mAP over the base model. Additionally, the GCEE module also enhances the performance by 1% mAP without the CSP-Decoder. As expected, our network achieves state-of-the-art results when both modules are combined, indicating that the combination of the two modules can extract cleaner features.

5.2. Loss Function

In our proposed ODFC-YOLO, we employ a multi-task learning paradigm to optimize the network by assigning different roles to the dehazing and detection tasks. We report the impact of different combinations of loss weights on detection performance in Table 7. Experimental results demonstrate that reducing the proportion of detection loss \mathcal{L}_{de} leads to an improvement in the network's performance. This is because the values of \mathcal{L}_{de} are much larger than those of the dehazing loss \mathcal{L}_{re} , which causes the network to be dominated by the detection subnet, preventing the dehazing subnet from producing clean features. By reducing the proportion of detection loss, the overall loss value is also reduced, helping the network converge. Additionally, the addition of noise ϵ improves the model's generalization ability. After conducting extensive experiments, we determined that the optimal loss weights are $\lambda_1 = 0.2$ and $\lambda_2 = 0.8$.

Table 7. Ablation study with different weight assignments for detection loss and dehazing loss. mAP* represents the detection performance applying the loss function without ϵ .

$\lambda_1 \& \lambda_2$	ϵ	mAP*	mAP
0.7&0.3	0.314	56.01	55.63
0.6&0.4	0.412	57.35	57.83
0.4&0.6	0.608	58.82	59.98
0.3&0.7	0.706	59.82	60.65
0.2&1.0	1.004	58.89	60.66
0.2&0.8	0.804	61.02	62.05

5.3. Inference Time

In our model, we introduce a multi-task joint learning approach to simultaneously localize and classify degraded images. Table 8 lists the average runtime and Frames Per Second (FPS) metrics for various advanced detection methods. It is observed that our method only needs 0.026 s to infer a hazy image. Despite the additional subnet and convolutional layers in our network, the dehazing subnet does not participate in the inference process, allowing our model to maintain a high detection speed while still achieving better performance than other methods.

Table 8. Detection speed comparison of the proposed ODFC-YOLO with different advance detection methods.

Method	Run Time	FPS
YOLOXs	0.018	55.6
DCP-YOLOXs	1.238	0.8
MS-DAYOLO	0.037	27.0
DS-Net	0.035	28.6
IA-YOLO	0.039	25.6
TogetherNet	0.031	35.1
ODFC-YOLO(Ours)	<u>0.026</u>	<u>36.5</u>

6. Conclusions

In this paper, we adopt a multi-task learning architecture that simultaneously performs image dehazing and object detection in foggy weather scenarios. By jointly optimizing these tasks, our approach can effectively improve object detection performance by leveraging the complementary nature of image dehazing and object detection. The proposed method effectively addresses the problem of object detection in foggy weather scenarios by using a Cross-Stage Partial Fusion Decoder (CSP-Decoder) to alleviate the discrepancy in feature region offset and the interference of foggy features on feature extraction, as well as a Global Context Enhancement Extraction (GCEE) module to extract additional information beneficial to detection from degraded images by building global feature long-range dependencies and enriching the content of output features. Furthermore, we design a Correlation-Aware Aggregated Loss to achieve optimal detection performance by averaging noise pattern and tuning gradient magnitudes between different tasks. Finally, through extensive experiments on synthetic and real-world fog datasets, our model surpasses the existing state-of-the-art method by 0.5 mAP on the RTTS dataset and by more than 1.6 mAP on the Foggy Driving dataset. Impressively, our model outperforms the previous best method by a remarkable 9.07 mAP points. Furthermore, our approach maintains real-time inference efficiency and achieves a commendable performance rate of 36.5 FPS. In future work, we plan to explore additional methods for improving detector performance in foggy weather, and apply our approach to real-world applications.

Author Contributions: Conceptualization, methodology, writing, funding acquisition, and supervision, G.Z. and Y.Z.; software, validation, and data curation, W.F. and Y.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Natural Science Foundation of China (Grant number: 62172231 and U20B2065); Natural Science Foundation of Jiangsu Province of China (Grant number: BK20211539 and BK20220107); the Jiangsu Provincial Postgraduate Practice Innovation Program (Grant number: SJCX23_0398); Chongqing Municipal Natural Science Foundation (Grant number: CSTB2022NSCQ-MSX0894).

Data Availability Statement: The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Grigorescu, S.; Trasnea, B.; Cocias, T.; Macesanu, G. A survey of deep learning techniques for autonomous driving. *J. Field Robot.* **2020**, *37*, 362–386. [\[CrossRef\]](#)
2. Zhang, G.; Ge, Y.; Dong, Z.; Wang, H.; Zheng, Y.; Chen, S. Deep high-resolution representation learning for cross-resolution person re-identification. *IEEE Trans. Image Process.* **2021**, *30*, 8913–8925. [\[CrossRef\]](#) [\[PubMed\]](#)
3. Zhang, G.; Liu, J.; Chen, Y.; Zheng, Y.; Zhang, H. Multi-biometric unified network for cloth-changing person re-identification. *IEEE Trans. Image Process.* **2023**, *32*, 4555–4566. [\[CrossRef\]](#) [\[PubMed\]](#)
4. Zhang, H.; Zhang, G.; Chen, Y.; Zheng, Y. Global Relation-Aware Contrast Learning for Unsupervised Person Re-Identification. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 8599–8610. [\[CrossRef\]](#)
5. Zhang, G.; Luo, Z.; Chen, Y.; Zheng, Y.; Lin, W. Illumination Unification for Person Re-Identification. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 6766–6777. [\[CrossRef\]](#)
6. Zhu, P.; Wen, L.; Du, D.; Bian, X.; Fan, H.; Hu, Q.; Ling, H. Detection and tracking meet drones challenge. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 7380–7399. [\[CrossRef\]](#) [\[PubMed\]](#)
7. Liu, F.; Zhang, X.; Wan, F.; Ji, X.; Ye, Q. Domain contrast for domain adaptive object detection. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 8227–8237. [\[CrossRef\]](#)
8. Cao, J.; Pang, Y.; Zhao, S.; Li, X. High-level semantic networks for multi-scale object detection. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *30*, 3372–3386. [\[CrossRef\]](#)
9. Zhang, S.; Wen, L.; Lei, Z.; Li, S.Z. RefineDet++: Single-shot refinement neural network for object detection. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *31*, 674–687. [\[CrossRef\]](#)
10. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
11. 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*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 740–755.
12. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. Yolox: Exceeding yolo series in 2021. *arXiv* **2021**, arXiv:2107.08430.
13. Zhang, G.; Fang, W.; Zheng, Y.; Wang, R. SDBAD-Net: A Spatial Dual-Branch Attention Dehazing Network based on Meta-Former Paradigm. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *1*. [\[CrossRef\]](#)
14. Li, B.; Ren, W.; Fu, D.; Tao, D.; Feng, D.; Zeng, W.; Wang, Z. Benchmarking single-image dehazing and beyond. *IEEE Trans. Image Process.* **2018**, *28*, 492–505. [\[CrossRef\]](#) [\[PubMed\]](#)
15. Chen, C.; Dou, Q.; Chen, H.; Qin, J.; Heng, P.A. Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Hilton, HI, USA, 27 January–1 February 2019; Volume 33, pp. 865–872.
16. 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*, Virtually, 22 February–1 March 2022; Volume 36, pp. 1792–1800.
17. 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 Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3339–3348.
18. 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*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 763–780.
19. Hnewa, M.; Radha, H. Multiscale domain adaptive yolo for cross-domain object detection. In *Proceedings of the 2021 IEEE International Conference on Image Processing (ICIP)*, Anchorage, AK, USA, 19–22 September 2021; IEEE: Manhattan, NY, USA, 2021; pp. 3323–3327.
20. Zhao, Z.Q.; Zheng, P.; Xu, S.T.; Wu, X. Object detection with deep learning: A review. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *30*, 3212–3232. [\[CrossRef\]](#)
21. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.

22. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
23. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*. [[CrossRef](#)] [[PubMed](#)]
24. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
25. Kala, J.R.; Kre, D.M.; Gnassou, A.N.; Kala, J.R.K.; Akpablin, Y.M.A.; Coulibaly, T. Assets management on electrical grid using Faster-RCNN. *Ann. Oper. Res.* **2020**, *308*, 307–320. [[CrossRef](#)]
26. Lu, X.; Li, B.; Yue, Y.; Li, Q.; Yan, J. Grid r-cnn. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7363–7372.
27. Uijlings, J.R.; Van De Sande, K.E.; Gevers, T.; Smeulders, A.W. Selective search for object recognition. *Int. J. Comput. Vis.* **2013**, *104*, 154–171. [[CrossRef](#)]
28. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
29. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 779–788.
30. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
31. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv* **2022**, arXiv:2207.02696.
32. Zhu, B.; Wang, J.; Jiang, Z.; Zong, F.; Liu, S.; Li, Z.; Sun, J. Autoassign: Differentiable label assignment for dense object detection. *arXiv* **2020**, arXiv:2007.03496.
33. Wang, Y.; Zhang, X.; Yang, T.; Sun, J. Anchor detr: Query design for transformer-based detector. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtually, 22 February–1 March 2022; Volume 36, pp. 2567–2575.
34. Li, F.; Zeng, A.; Liu, S.; Zhang, H.; Li, H.; Zhang, L.; Ni, L.M. Lite DETR: An interleaved multi-scale encoder for efficient detr. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 18558–18567.
35. Basak, S.; Rajendran, S.; Pollin, S.; Scheers, B. Combined RF-based drone detection and classification. *IEEE Trans. Cogn. Commun. Netw.* **2021**, *8*, 111–120. [[CrossRef](#)]
36. Anwar, M.Z.; Kaleem, Z.; Jamalipour, A. Machine learning inspired sound-based amateur drone detection for public safety applications. *IEEE Trans. Veh. Technol.* **2019**, *68*, 2526–2534. [[CrossRef](#)]
37. Huang, Y.; Chen, J.; Huang, D. UFPMP-Det: Toward accurate and efficient object detection on drone imagery. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtually, 22 February–1 March 2022; Volume 36, pp. 1026–1033.
38. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv* **2020**, arXiv:2010.04159.
39. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 213–229.
40. Ye, T.; Zhang, Y.; Jiang, M.; Chen, L.; Liu, Y.; Chen, S.; Chen, E. Perceiving and modeling density for image dehazing. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 130–145.
41. Guo, C.L.; Yan, Q.; Anwar, S.; Cong, R.; Ren, W.; Li, C. Image dehazing transformer with transmission-aware 3d position embedding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 5812–5820.
42. Bai, H.; Pan, J.; Xiang, X.; Tang, J. Self-guided image dehazing using progressive feature fusion. *IEEE Trans. Image Process.* **2022**, *31*, 1217–1229. [[CrossRef](#)] [[PubMed](#)]
43. Song, Y.; He, Z.; Qian, H.; Du, X. Vision transformers for single image dehazing. *IEEE Trans. Image Process.* **2023**, *32*, 1927–1941. [[CrossRef](#)] [[PubMed](#)]
44. Lu, L.; Xiong, Q.; Chu, D.; Xu, B. MixDehazeNet: Mix Structure Block For Image Dehazing Network. *arXiv* **2023**, arXiv:2305.17654.
45. Qiu, Y.; Lu, Y.; Wang, Y.; Jiang, H. IDOD-YOLOV7: Image-Dehazing YOLOV7 for Object Detection in Low-Light Foggy Traffic Environments. *Sensors* **2023**, *23*, 1347. [[CrossRef](#)]
46. Li, C.; Zhou, H.; Liu, Y.; Yang, C.; Xie, Y.; Li, Z.; Zhu, L. Detection-friendly dehazing: Object detection in real-world hazy scenes. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 8284–8295. [[CrossRef](#)]
47. Zhong, X.; Tu, S.; Ma, X.; Jiang, K.; Huang, W.; Wang, Z. Rainy WCity: A Real Rainfall Dataset with Diverse Conditions for Semantic Driving Scene Understanding. In Proceedings of the International Joint Conference on Artificial Intelligence, Vienna, Austria, 23–29 July 2022.
48. Hu, X.; Fu, C.W.; Zhu, L.; Heng, P.A. Depth-attentional features for single-image rain removal. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–19 June 2019; pp. 8022–8031.
49. Sakaridis, C.; Dai, D.; Van Gool, L. Semantic foggy scene understanding with synthetic data. *Int. J. Comput. Vis.* **2018**, *126*, 973–992. [[CrossRef](#)]
50. Huang, S.C.; Le, T.H.; Jaw, D.W. DSNet: Joint semantic learning for object detection in inclement weather conditions. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 2623–2633. [[CrossRef](#)]

51. Wang, Y.; Yan, X.; Zhang, K.; Gong, L.; Xie, H.; Wang, F.L.; Wei, M. TogetherNet: Bridging Image Restoration and Object Detection Together via Dynamic Enhancement Learning. *Comput. Graph. Forum* **2022**, *41*, 465–476. [[CrossRef](#)]
52. Li, J.; Zhou, P.; Chen, Y.; Zhao, J.; Roy, S.; Shuicheng, Y.; Feng, J.; Sim, T. Task relation networks. In Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa Village, HI, USA, 7–11 January 2019; IEEE: Manhattan, NY, USA, 2019; pp. 932–940.
53. Yu, J.; Jiang, Y.; Wang, Z.; Cao, Z.; Huang, T. Unitbox: An advanced object detection network. In Proceedings of the 24th ACM International Conference on Multimedia, Amsterdam, The Netherlands, 15–19 October 2016; pp. 516–520.
54. He, K.; Sun, J.; Tang, X. Single image haze removal using dark channel prior. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *33*, 2341–2353. [[PubMed](#)]
55. 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, Venice, Italy, 22–29 October 2017; pp. 4770–4778.
56. Chen, D.; He, M.; Fan, Q.; Liao, J.; Zhang, L.; Hou, D.; Yuan, L.; Hua, G. Gated context aggregation network for image dehazing and deraining. In Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa Village, HI, USA, 7–11 January 2019; IEEE: Manhattan, NY, USA, 2019; pp. 1375–1383.
57. 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, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 11908–11915.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.