Article

YOLOv5s-Fog: An Improved Model Based on YOLOv5s for Object Detection in Foggy Weather Scenarios

XiangLin Meng [1], Yi Liu [2], Lili Fan [2] and Jingjing Fan [1,2]*

[1] School of Electrical and Control Engineering, North China University of Technology, Beijing 100144, China
[2] National Industrial Innovation Center of Intelligent Equipment, Changzhou 213300, China
* Correspondence: jjfan@ncut.edu.cn

Abstract: In foggy weather scenarios, the scattering and absorption of light by water droplets and particulate matter cause object features in images to become blurred or lost, presenting a significant challenge for target detection in autonomous driving vehicles. To tackle this problem, this study proposes a foggy weather detection method, YOLOv5s-Fog, based on the YOLOv5s framework. The model enhances the feature extraction and expression capabilities of YOLOv5s by introducing a novel target detection layer, SwinFocus. Additionally, this research incorporates decoupled head into the model and replaces the conventional non-maximum suppression method with Soft-NMS. Experimental results demonstrate that these improvements effectively enhance the detection performance for blurry objects and small targets in foggy weather conditions. Compared to the baseline model YOLOv5s, YOLOv5s-Fog achieves a 5.4% increase in mAP on the RTTS dataset, reaching 73.4%. This method provides technical support for rapid and accurate target detection in adverse weather conditions, such as foggy weather, for autonomous driving vehicles.

Keywords: foggy weather scenarios; deep learning; SwinFoucs; decoupled head; Soft-NMS

1. Introduction

In the field of autonomous driving, object detection is a crucial technology [1], and its accuracy and robustness are of paramount importance for practical applications [2]. However, in foggy weather scenarios, challenges arise due to weakened light and issues such as blurred object edges, which lead to a decline in algorithm performance, consequently affecting the safety and reliability of autonomous vehicles [3]. Therefore, conducting research on target detection in foggy weather scenes holds significant significance.

In recent years, researchers have made certain progress in addressing the problem of object detection in foggy weather conditions [4,5]. Traditional methods primarily rely on conventional computer vision techniques such as edge detection, filtering, and background modeling. While these methods can partially handle foggy images, their effectiveness in complex scenes and under challenging foggy conditions is limited. To address the issue of object detection in complex foggy scenes, scholars have started exploring the utilization of physical models to represent foggy images. He et al. [6] proposed a single-image dehazing method based on the dark channel prior, while Zhu et al. [7] presented a fast single-image dehazing approach based on color attenuation prior. These dehazing methods improve the visibility of foggy images and subsequently enhance the accuracy of object detection. However, physical model-based methods require the estimation of fog density, making it difficult to handle multiple fog densities in complex scenes.

With the continuous development of deep learning techniques, deep learning has gradually become a research hotspot in the field of object detection [8,9]. Compared to traditional methods, deep learning models can directly learn tasks from raw data and exhibit improved generalization through training on large-scale datasets. Deep learning-based object detection algorithms can be categorized into two-stage detectors and one-stage detectors. Two-stage detectors first generate a set of candidate boxes and then perform

 doi:10.20944/preprints202305.0729.v1

classification and position regression for each candidate box. Faster R-CNN [10] is the most representative algorithm in this category, which employs an RPN [10] to generate candidate boxes and utilizes ROI Pooling [11] for classification and position regression of each candidate box. In addressing the problem of object detection in foggy weather conditions, Chen et al. [12] proposed a domain adaptive method that aligns features and adapts domains between source and target domains, thereby improving the detection performance in the target domain. However, region proposal-based methods require more computational resources and incur higher costs, making them less suitable for real-time applications with stringent timing requirements.

One-stage detectors directly perform classification and position regression on the input image without the need for generating candidate boxes. The most representative algorithms in this category are the YOLO series [13–15] and SSD [16]. YOLO divides the image into a grid and predicts bounding boxes and class probabilities for each grid cell, while SSD predicts bounding boxes of different sizes on different feature layers. Compared to two-stage detectors, one-stage detectors have a significant advantage in terms of speed, making them particularly suitable for real-time applications. However, improving the accuracy of object detection in complex weather and lighting conditions remains a challenge. Hnewa et al. [17] proposed a cross-domain object detection method that utilizes multi-scale features and domain adaptation techniques to enhance the detection performance in complex weather conditions. Liu et al. [18] designed a fully differentiable image processing module based on YOLOv3 [13] for object detection in foggy and low-light scenarios. Although this image-adaptive approach improves detection accuracy, it also introduces some undesirable noise.

In the aforementioned research on foggy weather object detection, although the detection accuracy has been improved, most of these methods are primarily focused on defogging and image enhancement [19]. This study aims to enable object detection algorithms to achieve clear detection in foggy weather scenes without any preprocessing of the original image. In recent years, the application of Transformer models [20–22] in computer vision has been increasing. These models leverage self-attention mechanisms to capture relationships within an image, thereby enhancing model performance. In this study, the Swin Transformer [22] component is incorporated into the YOLOv5s model to improve detection accuracy in adverse weather conditions.

The main contributions of this study are as follows:

1. On the basis of the YOLOv5s model, we introduce a multi-scale attention feature detection layer called SwinFocus, based on the Swin Transformer, to better capture the correlations among different regions in foggy images;
2. The traditional YOLO Head is replaced with a decoupled head, which decomposes the object detection task into different subtasks, reducing the model's reliance on specific regions in the input image;
3. In the stage of non-maximum suppression (NMS), Soft-NMS is employed to better preserve the target information, thereby effectively reducing issues such as false positives and false negatives.

The remaining sections of this paper are organized as follows. In Section 2, we provide a brief overview of the original YOLOv5s model and elaborate on the innovations proposed in this study. Section 3 presents the dataset, experimental details, and results obtained in our experiments. Finally, in Section 4, we summarize our work and propose some future research directions.

2. YOLOv5s-Fog

2.1. Overview of YOLOv5

YOLOv5 [15] is an efficient and highly accurate real-time object detection algorithm that extends the YOLO series [13,14]. This algorithm employs a single neural network to perform bounding box and category predictions. In comparison to its previous versions, YOLOv5 incorporates several improvements, including a new backbone network based

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 May 2023 doi:10.20944/preprints202305.0729.v1

on the CSP architecture [23], dynamic anchor allocation methods, and data augmentation techniques such as Mixup [24]. These enhancements have enabled the algorithm to achieve outstanding performance on multiple benchmark datasets while maintaining real-time inference speed on both CPU and GPU platforms. The YOLOv5 model consists of four different configurations: YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. In general, YOLOv5s is well-suited for real-time object detection in scenarios with limited computational resources, while YOLOv5x is more suitable for applications that require high-precision detection. Considering the real-time detection requirements in foggy weather conditions, this study employs YOLOv5s as the experimental model. The operational flow of YOLOv5s-Fog proposed in this paper is illustrated in Figure 1.

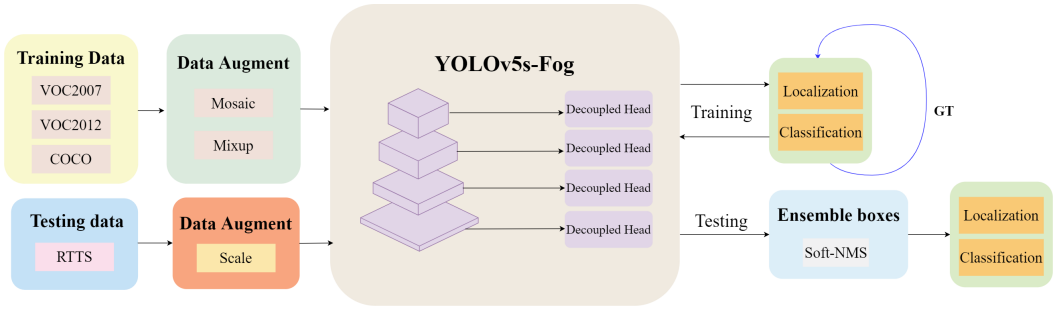


Figure 1. Operational Procedure of YOLOv5s-Fog.This framework incorporates an augmented predictive feature layer to bolster the network's regional comprehension. Additionally, we employ decoupled head to effectively address scenarios characterized by diminished contrast and indistinct boundaries. Lastly, the Soft-NMS technique is employed for the integration of bounding boxes.

2.2. Construction of Object Detection Model for Foggy Scenes

2.2.1. The Swin Transformer Architecture

Swin Transformer [22] is a neural network based on Transformer [20], which has shown excellent performance in computer vision tasks such as image classification, object detection, and semantic segmentation. The architecture of this network is illustrated in Figure 2. In Swin Transformer, the input image is first fed into the Patch Partition module for block-wise processing, where each 4×4 neighboring pixels are grouped into a patch and then flattened along the channel dimension. Subsequently, four stages are constructed to generate feature maps of different sizes. While Stage 1 employs a Linear Embedding layer as the initial step, the remaining three stages undergo Patch Merging for downsampling before being stacked together in a repetitive manner.

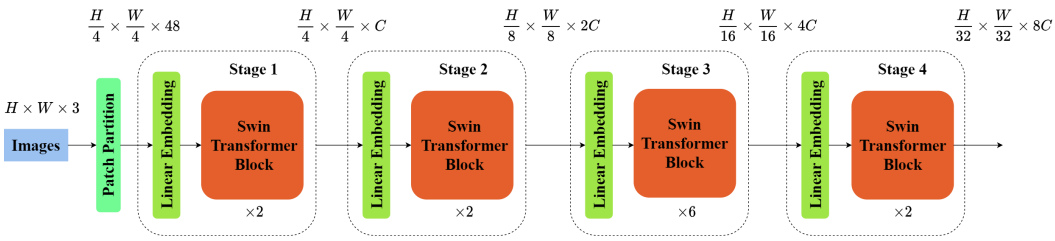


Figure 2. Swin Transformer Architecture.

The Swin Transformer Block is the fundamental building block in the Swin Transformer, and its structure is illustrated in Figure 3.

This module effectively addresses the issue of long-range dependencies and enables efficient processing of large-scale input data through techniques such as hierarchical feature representation, window-based attention mechanism, and depth-wise separable convolution. Specifically, the hierarchical feature representation utilizes multiple stages to downsample the input data and obtain feature representations at different resolutions. The window-based attention mechanism transforms the global attention mechanism into a local attention

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 May 2023 doi:10.20944/preprints202305.0729.v1

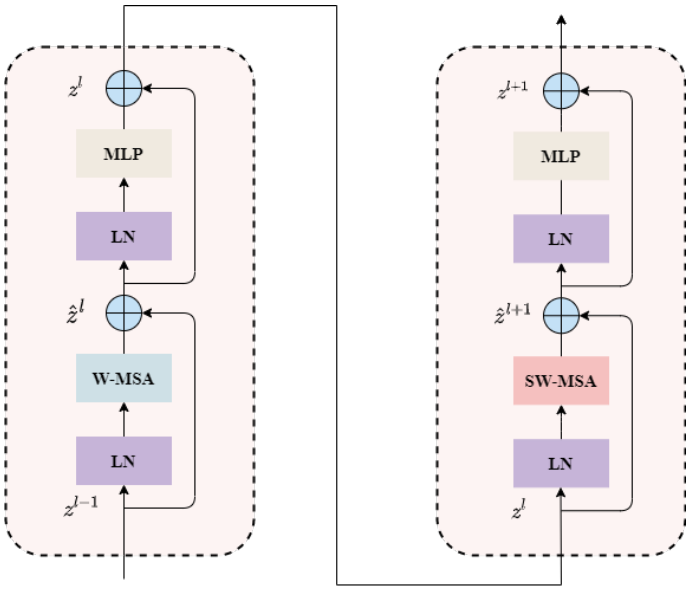


Figure 3. The consecutive Swin Transformer Blocks. The Swin Transformer Block first introduces Windows Multi-head Self Attention (W-MSA), which significantly reduces computational complexity compared to MSA. However, it only performs self-attention computations within each window. To enable information propagation between windows, the Shifted Windows Multi-Head Self-Attention (SW-MSA) is subsequently introduced.

mechanism, thereby reducing time and space complexity. The depth-wise separable convolution improves computational efficiency by reducing the number of parameters and computational complexity. The computation formulas for two consecutive Swin Transformer Blocks are as follows:

$$\hat{z}^l = W - MSA(LN(z^{l-1})) + z^{l-1} \tag{1}$$

$$z^l = MLP(LN(\hat{z}^l)) + \hat{z}^l \tag{2}$$

$$\hat{z}^{l+1} = SW - MSA(LN(\hat{z}^l)) + \hat{z}^l \tag{3}$$

$$\hat{z}^{l+1} = MLP(LN(\hat{z}^{l+1})) + \hat{z}^{l+1} \tag{4}$$

2.2.2. Decoupled Head

Decoupled Head [25,26] is a technique used for multi-task learning, which allows for shared feature representations between different tasks while training separate classifiers for each task. In object detection tasks, traditional methods use a shared feature detector, typically a convolutional neural network, to simultaneously predict both the class and location information for each object. However, this approach may face two challenges:

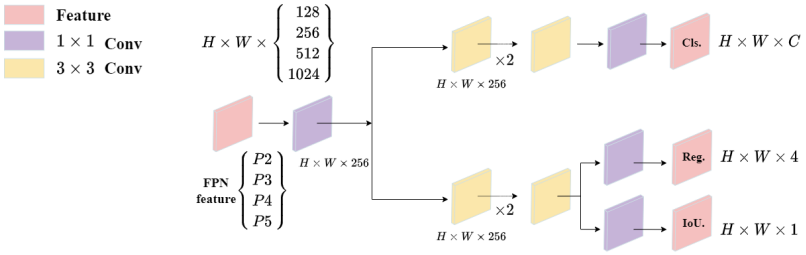


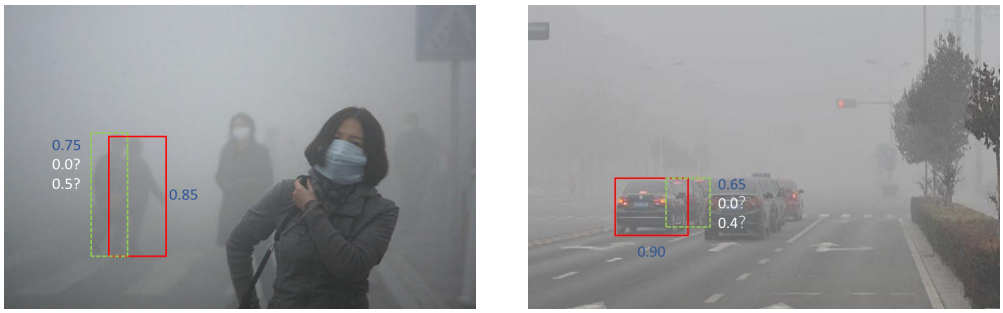
Figure 4. Decoupled Head Structure. Decoupled Head is a multi-task learning approach that divides object detection into two steps: image classification and object localization within the image.(YOLOv5s-Fog incorporates four detection heads).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 May 2023 doi:10.20944/preprints202305.0729.v1

firstly, the classification head and regression head are tightly coupled, requiring simultaneous adjustment of their loss functions during training, which can lead to complex and challenging training processes [27]; secondly, due to the varying sizes of different object types, a single detection head may struggle to adapt to all cases.The structure of the Decoupled Head, as shown in Figure 4, addresses these challenges by separating the classification head and regression head. Specifically, it utilizes a 1×1 convolutional layer to reduce the channel dimension, followed by two parallel branches, each consisting of two 3×3 convolutional layers [25]. This approach not only reduces the complexity of the network structure but also improves the model's accuracy.

2.2.3. Soft-NMS

Most object detection methods typically rely on post-processing using Non-Maximum Suppression (NMS). The conventional approach involves sorting the detection boxes based on their scores, keeping the box with the highest score, and removing other boxes with overlapping areas exceeding a certain threshold. However, this greedy algorithm exhibits the issue depicted in Figure 5: when two nearby boxes of the same class are present, a low-confidence box may be eliminated due to excessive overlap. Additionally, determining the threshold for NMS is not a trivial task. Soft-NMS (Soft Non-Maximum Suppression) [28]



(a) (b)

Figure 5. The issues that can occur during the post-processing stage of NMS. In Figure (a), there are two reliable pedestrian detections (green bounding box and red bounding box) with scores of 0.85 and 0.75, respectively. However, due to the significant overlap between the green and red bounding boxes, the green bounding box is assigned a lower score. The situation in Figure (b) is similar to that in Figure (a).

is a post-processing technique in object detection that improves the precision of detection results by considering the similarity between bounding boxes during the process of non-maximum suppression (NMS). Traditional NMS algorithms may suffer from issues such as excessive suppression or exclusion of correct bounding boxes, whereas Soft-NMS effectively addresses these problems.Its principle can be expressed by the following equation:For a set of input bounding boxes $B = \{b_1, b_2, ..., b_n\}$,where each bounding box b_i consists of four coordinates and a confidence score s_i, Soft-NMS measures their similarity by computing the Intersection over Union (IoU) values between the boxes:

$$IoU(b_i, b_j) = \frac{b_i \cap b_j}{b_i \cup b_j} \tag{5}$$

Then, the scores of each detection box are adjusted based on their similarity. Specifically, for the currently processed detection box b_i, its final weight is given by:

$$\omega_i{}^* = \begin{cases} s_i & \text{if } s_i > \theta \\ e^{-\frac{(IoU(b_i, b_k))^2}{\sigma}} \cdot s_k & \text{otherwise} \end{cases} \tag{6}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 May 2023 doi:10.20944/preprints202305.0729.v1

In this equation, θ represents a threshold. When s_i is greater than θ, the original score is retained. Otherwise, a Gaussian function is used to suppress other similar detection boxes, with σ controlling the rate of weight reduction. The final weight $\omega_i{}^*$ is adjusted based on a linear interpolation with the confidence score s_i of the current detection box.

$$\omega_i = (1 - \alpha)s_i + \alpha\omega_i{}^* \tag{7}$$

Among them, α is a parameter that controls the ratio between the adjusted score and the original score. Finally, for each detection box b_i, the Soft-NMS function adjusts it to ω_i, where detection boxes with higher similarity will appear with lower weights in the output results, thus avoiding issues such as excessive suppression and exclusion of correct detections.

2.3. The architecture of YOLOv5s-Fog network

The network architecture of YOLOv5s-Fog is shown in Figure 6. To address the

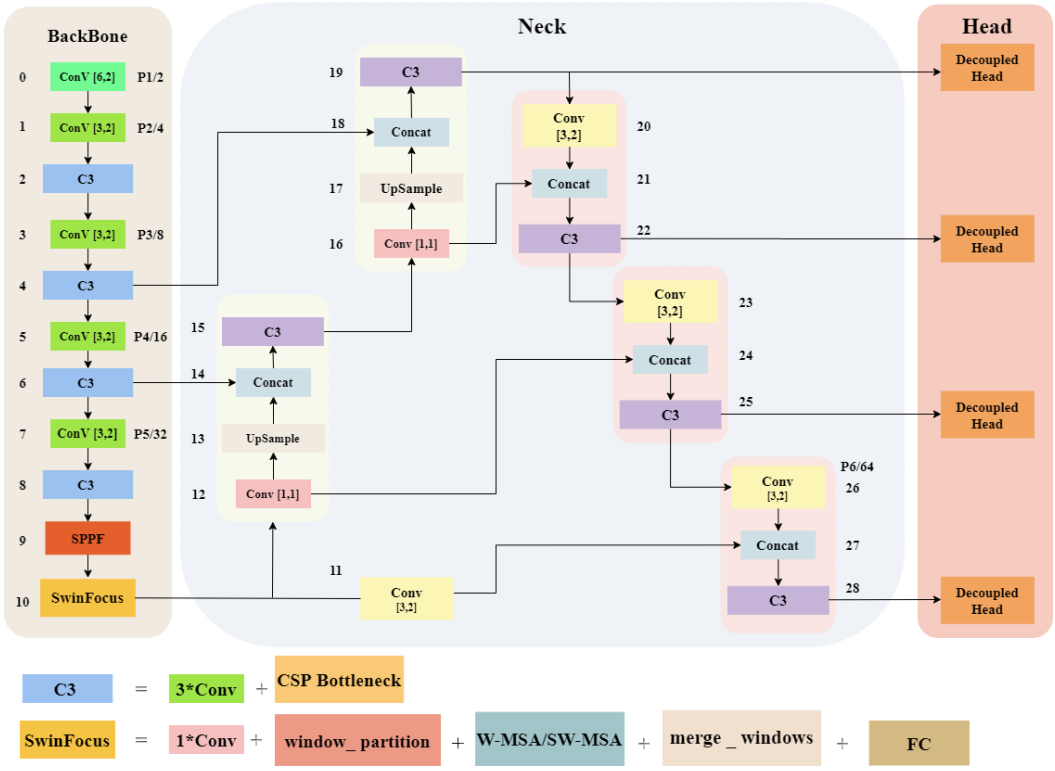


Figure 6. The network architecture of YOLOv5s-Fog introduces the following enhancements compared to the original version: (**a**) Addition of a target detection layer called SwinFocus based on Swin Transformer. (**b**) Use of a decoupled detection head to accomplish the final stage of the detection task.

challenges of target detection in low-visibility environments such as foggy weather, we have made improvements to the original YOLOv5s. Specifically, we have introduced an additional feature detection layer called SwinFocus, based on Swin Transformer, to mitigate the negative impact caused by variations in target scale due to weather conditions. SwinFocus decomposes the spatial and channel dimensions of the feature maps, allowing for global interaction and aggregation of information, enabling the network to better understand targets of different sizes and locations. Additionally, SwinFocus exhibits strong feature reuse capabilities, which enhance the model's ability to detect small objects in foggy conditions. In foggy scenes, limited illumination results in restricted details in the image, making it challenging to accurately extract object edges and texture features. To address this issue, we introduce Decoupled Head to replace the original detection head in YOLOv5. By utilizing shared representations, Decoupled Head encodes high-level semantic

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 May 2023 doi:10.20944/preprints202305.0729.v1

information shared across all tasks as feature representations. These shared features can better capture the shape, structure, and contextual information of objects in fog, thereby improving detection performance.

3. Experimental Setup and Results

3.1. Dataset

Insufficient datasets are available for training and testing object detection algorithms under adverse weather conditions, which can adversely affect their performance, particularly those based on CNN. Additionally, the traditional atmospheric scattering model [13] fails to accurately simulate real-world foggy scenes [18]. To ensure fairness, we selected a total of 8201 images as the training set (V_C_t), sourced from VOC [29] and COCO [30] datasets. For the test set, we utilized V_n_ts [18] and RTTS [31]. RTTS was employed to evaluate the method's object detection capability in foggy weather conditions, while V_n_ts was used to assess its performance on standard datasets. The dataset encompasses five categories: people, cars, buses, bicycles, and motorcycles. Further details regarding dataset usage are presented in Table 1.

Table 1. The relevant datasets used for training and testing purposes include V_C_t from VOC and COCO, V_n_ts from VOC2007_test, and RTTS, which is currently the only real-world foggy scene object detection dataset with multi-class detection labels.

Dataset	Image	Ps	Car	Bus	Bicycle	Motorcycle	Total
V_C_t	8201	14012	3471	850	1478	1277	21088
V_n_ts	2734	4528	337	1201	213	325	6604
RTTS	4322	7950	18413	1838	534	862	29597

3.2. Experimental Details

The experimental setup of YOLOv5s-Fog is shown in Table 2. During the training process of this study, we employed various effective data augmentation techniques, including MixUP [24] and Mosaic [14]. Additionally, we utilized a cosine learning rate scheduling strategy, setting the initial learning rate to 3e-4, batch size to 16, and conducting 30 iterations.

Table 2. Experimental Setup of YOLOv5s-Fog.

Configuration	Parameter
CPU	Intel Xeon(R) CPU E5-2678 v3
GPU	Nvidia Titan Xp*2
Pytorch	1.12
CUDA	11.1
cuDNN	8.5.0

3.3. Evaluation Metrics

This study evaluates the detection performance of the model using mean Average Precision (mAP). mAP is a metric commonly used to assess the performance of object detection algorithms. It represents the average area under the Precision-Recall curve, which provides a comprehensive evaluation of both the localization accuracy and recognition accuracy of the classifier. A higher mAP value indicates better detection performance of the model. The specific calculation method is as follows:

$$P = \frac{TP}{TP + FP} \tag{8}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 May 2023 doi:10.20944/preprints202305.0729.v1

$$R = \frac{TP}{TP + FN} \tag{9}$$

$$AP = \sum_{n} (R_n - R_{n-1}) P_n \tag{10}$$

$$mAP = \frac{1}{N} \sum_{i=1}^{N} AP_i \tag{11}$$

In this context,TP represents True Positive, FP represents False Positive, FN represents False Negative, R_n denotes Recall, P_n represents the maximum Precision at that Recall, and N indicates the number of classes.

3.4. Experimental Results

To validate the effectiveness of YOLOv5s-Fog, we compared it with various existing methods for foggy scene object detection, including deep learning-based object detection networks [13–15], dehazing methods [32,33], domain adaptation [17,34], and image adaptive enhancement [18]. The specific results are shown in Table 3.

Table 3. Comparison of the performance of each method on the conventional dataset (V_n_ts) and the foggy weather dataset (RTTS). The rightmost two columns present the mAP(%) on the two test datasets, including V_n_ts and RTTS.

Methods	V_n_ts	RTTS
YOLOv3 [13]	64.13	28.82
YOLOv3-SPP [35]	70.10	30.80
YOLOv4 [14]	79.84	35.15
MSBDN [32]	/	30.20
GridDehaze [33]	/	32.41
DAYOLO [17]	56.51	29.93
DSNet [34]	53.29	28.91
IA-YOLO [18]	72.65	36.73
YOLOv5 [15]	87.56	68.00
Ours	**92.23**	**73.40**

From Table 3, it can be observed that YOLOv5s-Fog outperforms other methods both on conventional weather datasets and foggy weather datasets. This is because YOLOv5s-

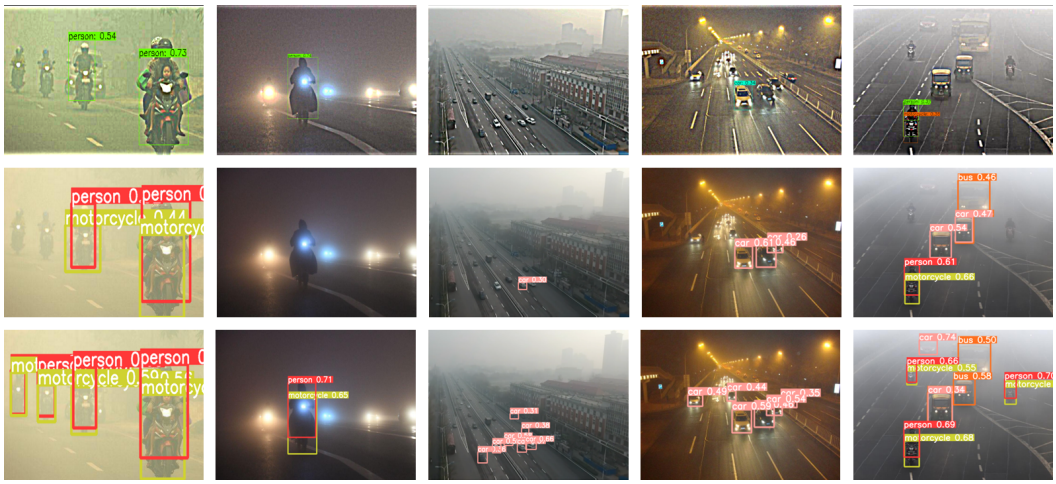


Figure 7. Partial detection results of IA-YOLO, YOLOv5s, and YOLOv5s-Fog on RTTS are shown below. The first row corresponds to IA-YOLO, the second row corresponds to YOLOv5s, and the third row corresponds to YOLOv5s-Fog.

Fog not only retains the high accuracy and fast speed of YOLOv5s itself but also incorporates a series of improvements specifically designed for foggy weather scenarios.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 May 2023 doi:10.20944/preprints202305.0729.v1

Compared to other deep learning-based object detection methods, YOLOv5s-Fog demonstrates stronger spatial awareness, effectively alleviating the challenges caused by object representation blurriness and significant scale variations due to weather conditions. Furthermore, YOLOv5s-Fog does not heavily focus on image dehazing, maintaining its original end-to-end detection approach and avoiding interference from artificially added noise during the detection phase. Figure 10 showcases partial detection results of the three models that performed well in RTTS. The first row presents IA-YOLO[18], which employs image adaptive techniques to remove specific weather information and restore the underlying content. Although this approach improves detection performance, it introduces undesired noise to the object detector. The second and third rows display the detection results of YOLOv5s and YOLOv5s-Fog, respectively, without image dehazing or image enhancement. It is evident from Figure 7 that YOLOv5s-Fog exhibits excellent detection capabilities in foggy weather conditions and low-light environments. Additionally, YOLOv5s-Fog can identify smaller objects in dense fog more effectively.

Figure 8 illustrates the loss curve of YOLOv5s-Fog during the training process and records the performance of each training stage on RTTS.

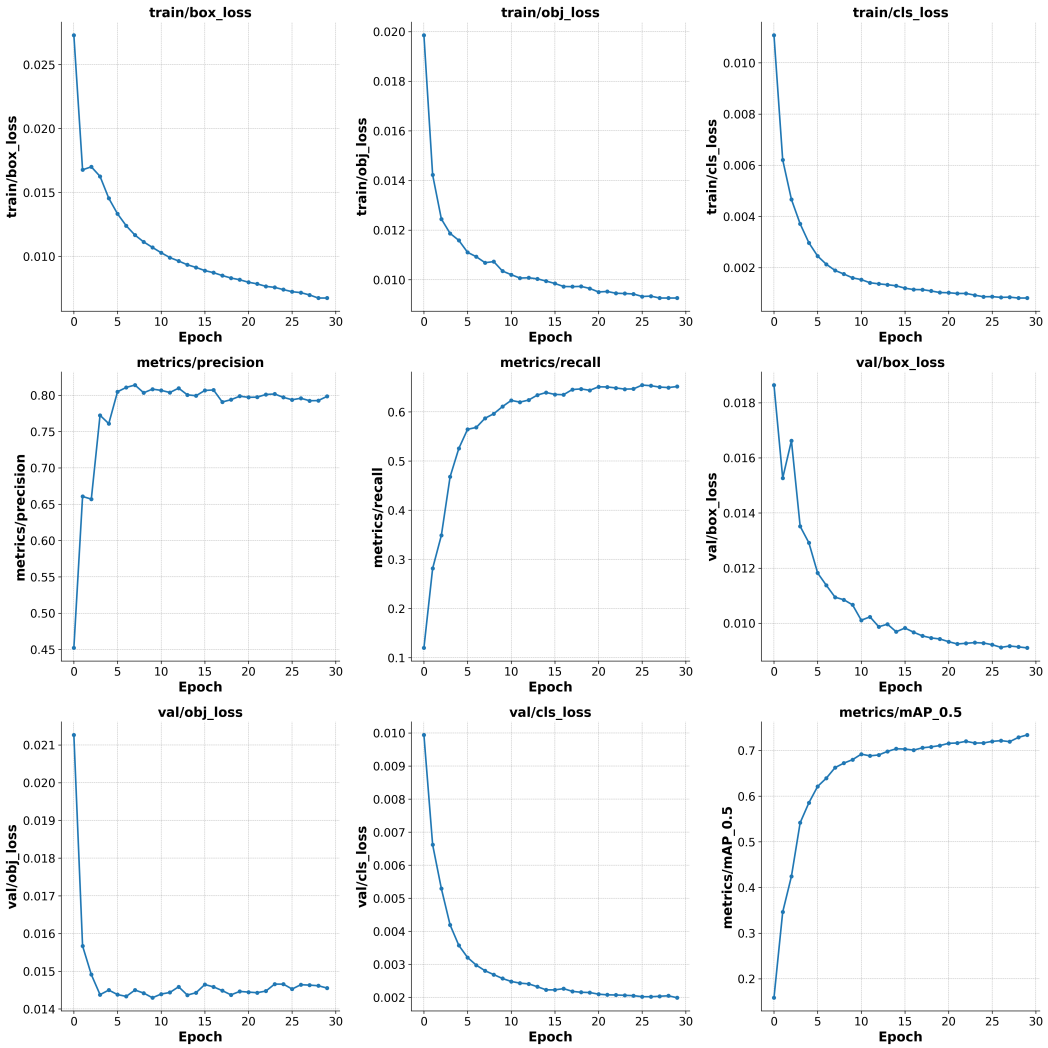


Figure 8. The loss curve during the training process of YOLOv5s-Fog and the performance of each training stage on RTTS (bottom right) are shown.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 May 2023 doi:10.20944/preprints202305.0729.v1

3.5. Ablation Studies

In order to validate the effectiveness of each module, we conducted an ablation study on the RTTS dataset. The impact of each module on the detection results is listed in Table 4.

Table 4. The Ablation Experiment on the RTTS Dataset.

Methods	**mAP** (%)	**mAP50-95** (%)	**GFLOPs**
YOLOv5s	68.00	41.17	15.8
YOLOv5s + SwinFocus	70.15 (↑2.15)	43.40 (↑2.23)	56.2
YOLOv5s + SwinFocus + Decoupled Head	71.79 (↑1.64)	44.38 (↑0.98)	57.4
YOLOv5s + SwinFocus + Decoupled Head + Soft-NMS	73.40 (↑1.61)	45.58 (↑1.20)	59.0

Table 5 documents the detection performance of YOLOv5s-Fog on each object category in the RTTS dataset after incorporating different modules

Table 5. The impact of incorporating the component on the precision (P) and recall (R) of the model was evaluated on the RTTS dataset.

Methods	**P**						**R**					
	All	Person	Car	Bus	Bicycle	Motorcycle	All	Person	Car	Bus	Bicycle	Motorcycle
YOLOv5s	0.87	0.912	0.926	0.795	0.86	0.856	0.489	0.725	0.504	0.318	0.485	0.413
YOLOv5s-Fog-1	0.74	0.69	0.911	0.753	0.647	0.7	0.635	0.641	0.632	0.496	0.697	0.712
YOLOv5s-Fog-2	0.88	0.924	0.938	0.835	0.83	0.88	0.55	0.735	0.51	0.397	0.614	0.493
YOLOv5s-Fog-3	0.78	0.851	0.762	0.675	0.81	0.807	0.70	0.809	0.793	0.601	0.694	0.631

The impact of the additional feature detection layer. Through experimental validation, we have observed that SwinFocus significantly enhances the model's mAP. This can be attributed to the adoption of the cross-domain self-attention mechanism during training, enabling the model to capture global features more effectively. Despite the introduction of additional object detection layers, which increase the model's parameters and computational burden, it is justified considering the application scenarios in adverse weather conditions like foggy weather. Table 4 demonstrates its notable performance improvements.
The impact of the decoupled head. By incorporating the Decoupled Head, the total number of layers in the model increased by 12, and the GFLOPs rose by 1.2. The adoption of the Decoupled Head not only enhances mAP but also enables adaptability to diverse object detection tasks and datasets, showcasing excellent scalability.
The impact of Soft-NMS. For object detection in foggy conditions, Soft-NMS primarily functions to address densely overlapping instances in large quantities. In Figure 9, we present the detection results of YOLOv5s-Fog on the RTTS dataset. Compared to traditional NMS, Soft-NMS exhibits superior handling of similar objects in complex environments, highlighting its significant advantage.

4. Conclusions

In this paper, we propose YOLOv5s-Fog, a novel approach to address the challenges of object detection under foggy conditions. Unlike previous research, we do not rely on dehazing or adaptive enhancement techniques applied to the original images. Instead, we enhance the YOLOv5s model by introducing additional detection layers and integrating advanced modules. Our improved model demonstrates higher accuracy in foggy conditions. Experimental results show the potential of our proposed method in object detection tasks under adverse weather conditions. In the future, we plan to invest more efforts in constructing datasets for object detection in extreme weather conditions and develop more efficient network architectures to enhance the model's accuracy in extreme weather detection.

 doi:10.20944/preprints202305.0729.v1

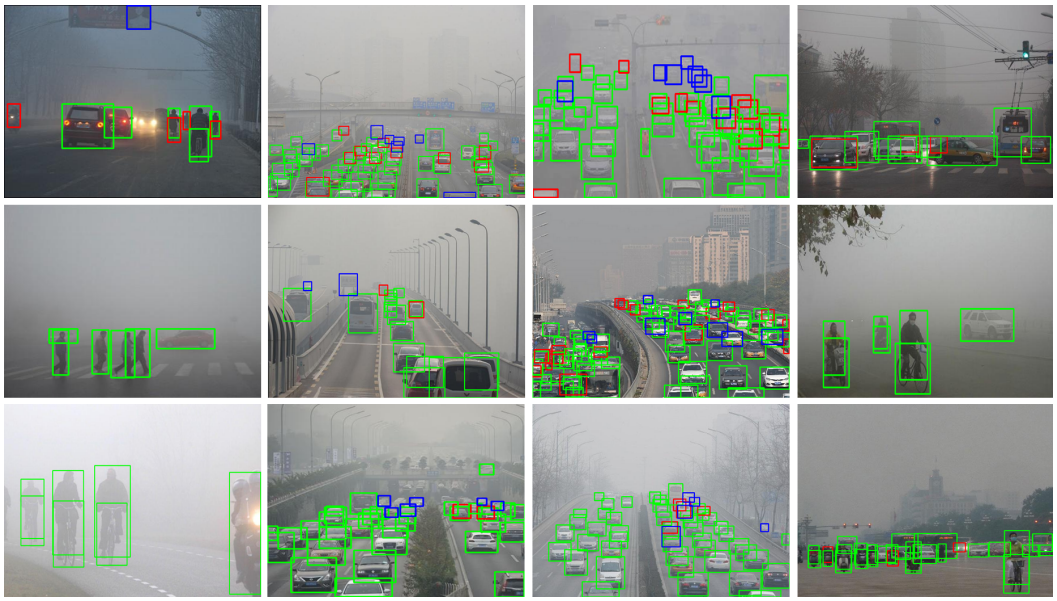


Figure 9. Visualization of the detection results of YOLOv5s-Fog on the RTTS dataset. The green, blue, and red boxes represent true positive (TP), false positive (FP), and false negative (FN) detections, respectively.

Author Contributions: Methodology, L.F.; software X.M.and Y.L.; validation J.F. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the Academician Mao Ming Workstation and the National Industrial Innovation Center of Intelligent Equipment(XS-JSFW-KCZNJS-202303-001).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: The authors would like to acknowledge the anonymous reviewers and editors whose thoughtful comments helped to improve this manuscript.

Conflicts of Interest: There are no conflicts of interest associated with the publication of this paper.

References

1. Bijelic, M.; Gruber, T.; Mannan, F.; Kraus, F.; Ritter, W.; Dietmayer, K.; Heide, F. Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11682–11692.
2. Walambe, R.; Marathe, A.; Kotecha, K.; Ghinea, G.; et al. Lightweight object detection ensemble framework for autonomous vehicles in challenging weather conditions. *Computational Intelligence and Neuroscience* **2021**, *2021*.
3. Liu, Z.; He, Y.; Wang, C.; Song, R. Analysis of the influence of foggy weather environment on the detection effect of machine vision obstacles. *Sensors* **2020**, *20*, 349.
4. Hahner, M.; Sakaridis, C.; Dai, D.; Van Gool, L. Fog simulation on real LiDAR point clouds for 3D object detection in adverse weather. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 15283–15292.
5. Krišto, M.; Ivasic-Kos, M.; Pobar, M. Thermal object detection in difficult weather conditions using YOLO. *IEEE access* **2020**, *8*, 125459–125476.
6. He, K.; Sun, J.; Tang, X. Single image haze removal using dark channel prior. *IEEE transactions on pattern analysis and machine intelligence* **2010**, *33*, 2341–2353.
7. Zhu, Q.; Mai, J.; Shao, L. A fast single image haze removal algorithm using color attenuation prior. *IEEE transactions on image processing* **2015**, *24*, 3522–3533.

8. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 580–587.
9. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980–2988.
10. 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* **2015**, *28*.
11. Girshick, R. Fast r-cnn. In Proceedings of the Proceedings of the IEEE international conference on computer vision, 2015, pp. 1440–1448.
12. 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 Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 3339–3348.
13. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779–788.
14. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934* **2020**.
15. Jocher, G.; Stoken, A.; Borovec, J.; Chaurasia, A.; Changyu, L.; Hogan, A.; Hajek, J.; Diaconu, L.; Kwon, Y.; Defretin, Y.; et al. ultralytics/yolov5: v5. 0-YOLOv5-P6 1280 models, AWS, Supervise. ly and YouTube integrations. *Zenodo* **2021**.
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 Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer, 2016, pp. 21–37.
17. 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). IEEE, 2021, pp. 3323–3327.
18. Liu, W.; Ren, G.; Yu, R.; Guo, S.; Zhu, J.; Zhang, L. Image-Adaptive YOLO for Object Detection in Adverse Weather Conditions. *Proceedings of the AAAI Conference on Artificial Intelligence* **2022**, p. 1792–1800. https://doi.org/10.1609/aaai.v36i2.20072.
19. Li, B.; Peng, X.; Wang, Z.; Xu, J.; Feng, D. Aod-net: All-in-one dehazing network. In Proceedings of the Proceedings of the IEEE international conference on computer vision, 2017, pp. 4770–4778.
20. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Advances in neural information processing systems* **2017**, *30*.
21. 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**.
22. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 10012–10022.
23. 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 Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, 2020, pp. 390–391.
24. Zhang, H.; Cisse, M.; Dauphin, Y.N.; Lopez-Paz, D. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412* **2017**.
25. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430* **2021**.
26. Tian, Z.; Shen, C.; Chen, H.; He, T. Fcos: Fully convolutional one-stage object detection. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 9627–9636.
27. Song, G.; Liu, Y.; Wang, X. Revisiting the sibling head in object detector. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 11563–11572.
28. Bodla, N.; Singh, B.; Chellappa, R.; Davis, L.S. Soft-NMS–improving object detection with one line of code. In Proceedings of the Proceedings of the IEEE international conference on computer vision, 2017, pp. 5561–5569.
29. Everingham, M.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The pascal visual object classes (voc) challenge. *International journal of computer vision* **2010**, *88*, 303–338.

 doi:10.20944/preprints202305.0729.v1

30. 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 Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer, 2014, pp. 740–755.
31. 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* **2018**, *28*, 492–505.
32. 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 Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 2157–2167.
33. Liu, X.; Ma, Y.; Shi, Z.; Chen, J. Griddehazenet: Attention-based multi-scale network for image dehazing. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 7314–7323.
34. 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* **2020**, *43*, 2623–2633.
35. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2015**, p. 1904–1916. https://doi.org/10.1109/tpami.2015.2389824.