Feature Adaptive YOLO for Remote Sensing Detection in Adverse Weather Conditions

Chaojun Ni, Wenhui Jiang*, Chao Cai, Qishou Zhu, Yuming Fang
School of Information Management,
Jiangxi University of Finance and Economics, Nanchang, China

***Abstract*—Target detection in remote sensing has been one of the most challenging tasks in the past few decades. However, the detection performance in adverse weather conditions still needs to be satisfactory, mainly caused by the low-quality image features and the fuzzy boundary information. This work proposes a novel framework called Feature Adaptive YOLO (FA-YOLO). Specifically, we present a Hierarchical Feature Enhancement Module (HFEM), which adaptively performs feature-level enhancement to tackle the adverse impacts of different weather conditions. Then, we propose an Adaptive receptive Field enhancement Module (AFM) that dynamically adjusts the receptive field of the features and thus can enrich the context information for feature augmentation. In addition, we introduce Deformable Gated Head (DG-Head) which reduces the clutter caused by adverse weather. Experimental results on RTTS and two synthetic datasets demonstrate that our proposed FA-YOLO significantly outperforms other state-of-the-art target detection models.**

***Index Terms*—Target detection, Remote sensing, Adverse weather, Feature fusion**

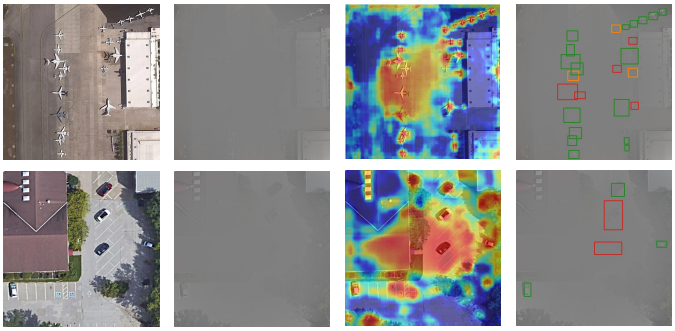

Fig. 1. Failure detection examples from the YOLOV5 model. (a) Clean images. (b) The same image under adverse weather conditions. (c) Activations of feature maps from YOLOV5. (d) Detection results, where green boxes indicate correct detections, red boxes indicate false detections, orange boxes indicate missed detections.

I. Introduction

Object detection in remote sensing images is tough due to expansive coverage and complex backgrounds with phenomena like clouds, haze, and droplets [1]–[3]. These elements reduce recognition accuracy. Therefore, the effectiveness of contemporary target detection methods still requires enhancement in adverse weather conditions [4]–[10].

The poor performances of existing state-of-the-art methods for detection in adverse weather can attribute to two prominent challenges: 1) Fuzzy feature information resulting from dense fog blurring context and impeding feature representation ability, which is further impaired by the incorporation of weather-induced clutter in the fusion process [11], [12]. 2) Insufficient boundary information due to fog, which leads to a mix of target edge information and weather-related clutter, complicating object boundary determination. As shown in Fig. 1(c), in thick fog, YOLOV5 [13] cannot focus on the object properly due to fuzzy features and the unclear boundary between objects.

An efficient approach to address the above problems is to perform image enhancement before object detection. For example, MSBDN [14] developed a boosted decoder to progressively restore the haze-free image by incorporating the Strengthen-Operate-Subtract boosting strategy. DSNet [15] proposed two sub-networks for the joint learning of image enhancement and object detection to deal with degenerated features' representation by sharing the feature extraction layer. IA-YOLO [16] presented a DIP module where a convolutional neural network predicts the parameters for feature enhancement and alleviates the problem of fuzzy feature information. Yang et al. [17] combined AdaInt with CNN predictions to achieve a more flexible sampling point allocation by adaptively learning the non-uniform sampling intervals in the 3D color space. However, in these works, image enhancement is not optimized for object detectors since latent information useful for detection may be neglected, and the clutter and lack of boundary information caused by adverse weather are ignored.

In this work, we introduce a novel end-to-end remote sensing detection network. It mainly includes three modules, a Hierarchical Feature Enhancement Module (HFEM), an Adaptive receptive Field enhancement Module (AFM), and a Deformable Gated Head (DG-Head). The HFEM module is designed to stratify the image to get features first, make them controlled by multiple filters, and perform the fusion, which can process defogging and cloud removal adaptively, making features clean. The AFM structure, designed to adaptively adjust the receptive field size and enrich context information according to the original features, processes multi-scale features generated when the enhanced feature is input to the backbone. Adding boundary information not only aids in category judgment but also improves the accuracy of the prediction

This work was supported in part by the National Natural Science Foundation of China (No.62132006 and No.62161013), and in part by the Natural Science Foundation of Jiangxi Province (No.20223AEI91002 and No.20224BAB212010).

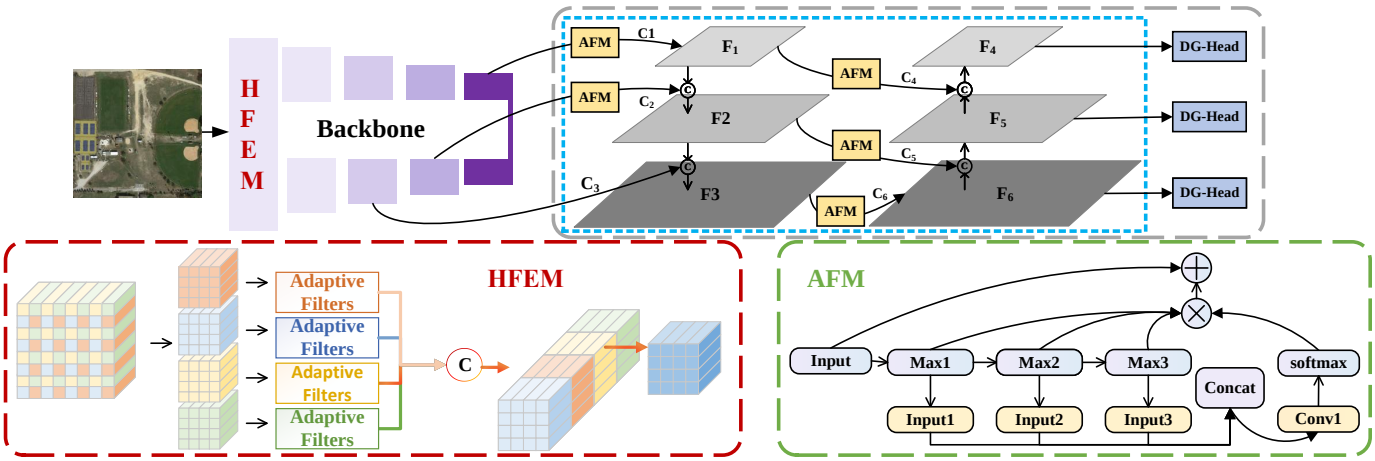

Fig. 2. Algorithm framework diagram of FA-YOLO.

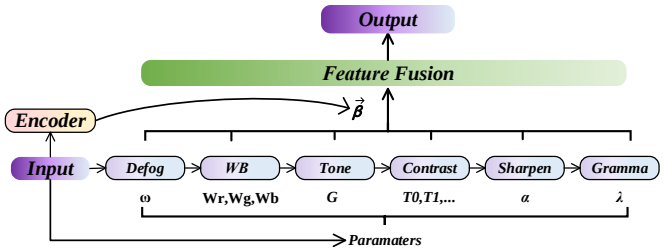

Fig. 3. The structure of Adaptive Filters.

bounding box in complex background environments. Finally, we propose DG-Head, which introduces deformable convolution and gating mechanisms to remove the clutter caused by adverse weather, refine the prediction box accurately, and prevent key features from being submerged in the clutter.

FA-YOLO outperforms the best competitor by 7.9%, 5.5%, and 4.9 % in terms of average detection precision on Dior_Foggy, Dior_Severe_Foggy and RTTS, respectively, demonstrating that our method effectively improves object detection performance in adverse weather conditions.

II. Proposed Method

As shown in Fig. 2, at the forefront of the backbone, HFEM stratifies the image and performs filtering and fusion. The feature fusion process introduces the AFM, dynamically adjusting the receptive field size and enriching context information. Finally, the DG-Head effectively removes clutter caused by adverse weather conditions and accurately refines prediction boxes.

A. Hierarchical Feature Enhancement Module

We propose HFEM to adapt features, including multiple filters [16] (Defog, WB, Tone, Contrast, Sharpen, Gramma filter) operations performed concurrently. At the same time, we add hierarchical and adaptive fusion mechanisms.

To achieve hierarchical control, we split the input feature map into four maps by taking a value at every other pixel in a feature. This gives us four complementary features of the same size without losing information.

The parameter prediction part (Encoder in Fig. 3) mainly predicts the parameters of filters and the parameters $\vec{\beta}$ of the fusion of each filter output. We use five convolution blocks, a global average pooling, and four fully connected layers.

Each filter takes the features from the previous layer as input and generates an enhanced feature. This enhanced feature is then multiplied by the corresponding β_i parameter and serves as the input for feature fusion. We label the outputs of the filters as F_1 to F_6, with each corresponding to a unique β_i value. Finally, all the enhanced features are fused to obtain the adaptive feature. The fusion operation can be expressed using the following equation:

$$\text{Output} = \sum_{i=1}^{6} \beta_i \cdot F_i \tag{1}$$

B. Adaptive Receptive Field Enhancement Module

We obtain the context information of different receptive fields by performing Max-Pooling with different kernel sizes. The kernel size is 3×3, 9×9, and 13×13. Then, we used an adaptive fusion approach. We assume the input size is (bs, C, H, W). We can obtain a spatial adaptive weight of $(bs, 3, H, W)$ after Max-Pooling, convolution, and SoftMax operation. The three channels correspond one-to-one to the three inputs, and calculating the weighted sum can aggregate the context information in the output adaptively.

The mode inspires AFM that humans recognize objects. For example, when some shadows in a thick cloud in a very high sky, if you restrict your vision to a thick cloud, you can judge that there is an object inside, but can't judge that it is a bird flying in the thick cloud or a person running in the dense fog land. But if you no longer limit your vision to the clouds and confirm that the background is the sky. The former is much more likely.

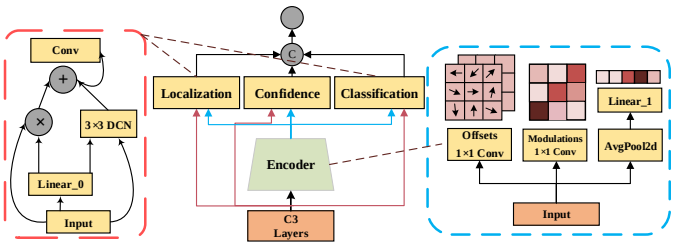

Fig. 4. The structure of DG-Head.

C. Deformable-Gated Head

DG-Head uses deformable convolution to predict the offsets of the convolution, introduces the modulation mechanism to reduce cluster, and dynamically adjusts the feature extraction range of the convolution based on the weather condition and the feature of the object. In the blue part of Fig. 4, we apply convolution to obtain offsets, modulation needed by deformable convolution, and gating scalars named α. Each processing part contains a linear layer, a deformable convolution, and a standard convolution. The linear layer (the module named Liner_0) computes two scalars called α_{org} and α_{def} by α and then input the output to the tanh function (which ranges from 0 to 1). The output of each task is the weighted sum of the output of the deformable convolution times α_{org} and the original feature times α_{def}. The obtained adjusted features are then fed into the final ordinary convolutional layer. Then the tensors obtained from the three tasks are decoded.

III. EXPERIMENTS

A. Datasets

We evaluate our proposed method on the RTTS and the remote sensing image dataset in adverse weather.

RTTS: The RTTS [18] dataset contains 4,322 genuine foggy images and five annotated object classes, specifically, person, bicycle, car, bus, and motorcycle.

Dior: The Dior [19] is a large-scale, publicly available benchmark for object detection in optical remote sensing images.

Dior_Foggy: We build a foggy dataset based on the Dior dataset according to the atmospheric scattering model (ASM) [20] as same as IA-YOLO. We employ the ASM to generate six different levels of fog. To help our model learn fog-invariant features, we employ a hybrid strategy using a mix of foggy and clear images (in a 2:1 ratio).

Dior_Severe_Foggy: To verify the performance of our method in object detection when various complex occlusions overlap, we randomly select 1 to 3 centers on each image and choose different fog concentrations to simulate the case of multiple visible polymers occlusion each other. At the same time, we use the same strategy as for Dior_Foggy.

B. Implementation Details

To ensure the fairness of comparison, all experiments are conducted with PyTorch. All experiments are trained by resizing images to 640×640 and with a batch size of 16 for 100 epochs. We apply a cosine learning rate scheduler with learning rates ranging from 1×10^{-5} to 1×10^{-4}, and the weight decay is 3×10^{-5}.

Fig. 5. Synthetic datasets of remote sensing in adverse conditions. The first row is Dior_Foggy, the second row is Dior_Severe_Foggy.

TABLE I
ABLATION EXPERIMENTS OF THE HFEM.

Hierarchical	Adaptive fusion	AP_{50}	AP
		62.9	40.2
✓		64.6	41.8
	✓	65.4	42.5
✓	✓	**66.8**	**43.2**

C. Ablation Studies

We design ablation experiments in different settings to verify each module's effectiveness and contribution rate.

HFEM: Our proposed Hierarchical and adaptive fusion mechanism helps fully utilize each filter output. To illustrate the effect, we conducted ablation experiments on HFEM, as seen in Table I. Without the adaptive fusion, the performance reduces by 2.3% AP in the Dior_Foggy dataset (compare the first and third rows of Table I). As we can see in Table II when we add the HFEM to YOLOV5, the AP increases by 3.0% on the Dior_Foggy dataset.

AFM: We discuss the effect of AFM at different positions from C1 to C6 (marked in Fig. 2). The results in Table II show a decrease in AP when increasing C3 alone. This is because C2 features are in a shallow position, which has more noise chaos. The AP increases when the AFM structure is added alone at other positions. When we add AFMs to these positions, our results are 2.2% higher than the baseline on the AP (compare the first and third rows of Table II).

DG-Head: The primary purpose of the DG-Head is to refine the prediction box and reduce clutter. According to the results in the fourth row of the Table III, the DG-Head improves both AP_{50} and AP by 4.1% and 3.2%.

D. Comparison with State-of-the-arts

Quantitative Analysis: We compare our model with several recent methods. The methods we compare include Faster RCNN [4], and DETR [6], which are common target detection methods, and DSNet [15], DAYOLO [22], and IA-YOLO [12], which are designed specifically for target detection in

TABLE II
ABLATION EXPERIMENTS OF THE AFM. FROM C1 TO C6 REPRESENT THE DIFFERENT CONCAT POSITIONS MARKED IN FIG. 2.

C1	C2	C3	AP_{50}	C4	C5	C6	AP_{50}
			62.9				62.9
✓			64.3	✓			63.8
	✓		63.5		✓		63.7
		✓	62.5			✓	63.2
✓		✓	62.2	✓		✓	64.2
	✓	✓	63.3		✓	✓	64.5
✓	✓		**64.9**	✓	✓		63.7
✓	✓	✓	64.3	✓	✓	✓	**64.8**

TABLE III
ABLATION EXPERIMENTS ON THE DIOR_FOGGY.

HFEM	AFM	DG-Head	P	R	AP_{50}	AP
			82.5	61.2	62.9	40.2
✓			85.2	62.3	66.3	43.2
	✓		84.6	63.4	65.3	42.4
		✓	85.6	63.6	67.0	43.4
✓	✓		87.2	63.6	67.2	45.3
✓		✓	88.2	64.0	67.7	45.8
	✓	✓	89.2	65.2	68.2	46.3
✓	✓	✓	**90.1**	**65.7**	**70.3**	**47.8**

TABLE IV
PERFORMANCE COMPARISONS WITH STATE-OF-THE-ART METHODS ON THE DIOR_FOGGY AND DIOR_SEVERE_FOGGY.

Algorithm	Backbone	Dior_Foggy		Dior_Severe_Foggy	
		AP_{50}	AP	AP_{50}	AP
Faster RCNN [8]	ResNet-101-FPN	56.9	33.5	49.7	31.8
RetinaNet [21]	ResNet-101-FPN	60.3	38.3	51.5	33.3
DETR [6]	ResNet-101-FPN	59.5	37.2	53.6	34.3
CENTERNET [7]	ResNet-101-FPN	57.1	36.1	49.0	30.6
MSBDN [14]	-	54.7	32.1	47.3	28.1
DSNet [15]	-	57.8	36.9	48.6	29.8
DAYOLO [22]	CSP-Darknet53	58.3	36.3	51.0	32.7
YOLOV5S [13]	CSP-Darknet53	62.9	40.2	53.5	33.7
AOD-YOLOV5S [23]	CSP-Darknet53	61.4	42.1	54.3	32.6
DehazeNet-YOLOV5S [24]	CSP-Darknet53	60.8	41.3	53.5	35.2
FFA-YOLOV5S [25]	CSP-Darknet53	63.2	42.3	54.5	34.7
IA-YOLO [12]	CSP-Darknet53	61.7	39.9	55.8	34.5
Ours	CSP-Darknet53	**70.3**	**47.8**	**60.8**	**40.0**

adverse weather. Meanwhile, five prominent dehazing approaches, including AOD-Net [23], DehazeNet [24], FFA-Net [25], are combined with YOLOV5S to create five hybrid models: AOD-YOLOV5S, DehazeNet-YOLOV5S, and FFA-YOLOV5S. All dehazing algorithms are trained using the complete ITS (Indoor Training Set) dataset [26], following the parameters outlined in their respective publications. The results of comparative experiments on the Dior_Foggy and Dior_Severe_Foggy datasets are listed in Table IV. The FA-YOLO achieves higher AP_{50} than the other frameworks in all the datasets and metrics. In particular, the FA-YOLO outperforms the IA-YOLO by 7.9% AP in Dior_Foggy, demonstrating our improvements further making up for the deficiencies of IA-YOLO. In the more challenging tasks, FA-YOLO is still 5.5% ahead of IA-YOLO in AP, which shows that it can be well applied in multiple occluded scenes. Furthermore, we

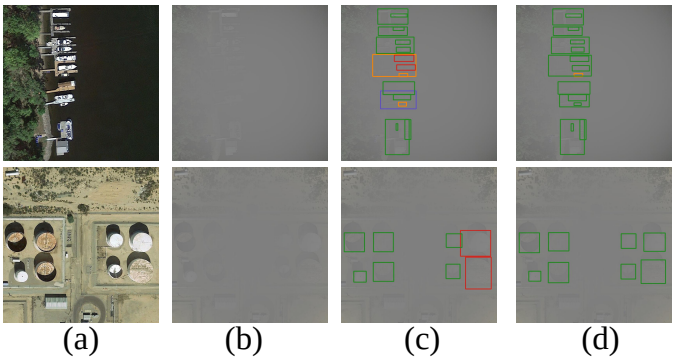

Fig. 6. Visualization of detection results. (a) Clean images. (b) The same image under adverse weather conditions. (c) Results of IA-YOLO. (d) Results of FA-YOLO, where green boxes indicate correct detections, red boxes indicate false detections, and orange boxes indicate missed detections.

TABLE V
PERFORMANCE COMPARISONS WITH STATE-OF-THE-ART METHODS ON THE RTTS

Method	Person	Bicycle	Car	Motorbike	Bus	AP_{50}
Faster RCNN [8]	74.3	45.4	60.6	38.4	24.8	48.7
RetinaNet [21]	74.5	52.1	60.4	37.4	21.7	49.2
DETR [6]	74.6	46.3	57.5	32.5	20.1	46.2
CENTERNET [7]	73.3	43.2	61.8	37.4	25.8	48.3
MSBDN [14]	74.4	45.8	60.4	36.8	23.5	48.2
DSNet [15]	75.5	42.3	61.2	40.5	24.2	48.7
DAYOLO [22]	73.8	47.3	58.2	30.2	21.2	46.1
YOLOV5S [13]	75.9	48.2	62.9	40.1	23.8	50.2
AOD-YOLOV5S [23]	76.6	43.4	60.9	34.7	22.1	47.5
DehazeNet-YOLOV5S [24]	75.9	46.8	62.8	40.3	24.7	50.1
FFA-YOLOV5S [25]	76.4	48.2	64.2	39.8	23.8	50.5
IA-YOLO [12]	72.2	41.1	60.2	37.8	21.3	46.5
GDIP-YOLO [27]	73.4	43.4	63.1	39.4	23.2	48.5
Togethernet [28]	76.5	55.3	68.5	44.2	29.7	54.8
Ours	**76.8**	**58.6**	**72.4**	**53.6**	**37.2**	**59.7**

conducted experiments on the real dataset RTTS, and it can be observed that our method achieved excellent results in real-world scenarios, surpassing the performance of the current best method by 4.9% in AP_{50}.

Qualitative Analysis: We compare the results of FA-YOLO with IA-Yolo, shown in Fig. 6 on our dataset. Our method can detect objects in extremely foggy conditions, which are generally missed by IA-Yolo (see the first row of Fig. 6). And, from the second row of Fig. 6, we can see that FA-YOLO's bounding box is more accurate. At the same time, IA-YOLO has obvious deviation in adverse weather.

IV. CONCLUSION

In this paper, aimed to address the problems caused by adverse weather, we propose a YOLOv5-based detection called FA-YOLO. For the issues of fuzzy feature information, HFEM is designed to suppress weather interference and restore latent details. AFM is proposed to enrich the context information for the problems of insufficient boundary information by adapting the receptive field size. DG-Head introduces deformable convolution to reduce clutter caused by combining multi-scale fusion and AFM structure to refine the bounding box. Finally, we achieve consistent improvements over state-of-the-art methods on RTTS and two synthetic datasets.

REFERENCES

[1] L. Dong and J. Shan, "A comprehensive review of earthquake-induced building damage detection with remote sensing techniques," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 84, pp. 85–99, 2013.

[2] Z. Deng, H. Sun, S. Zhou, J. Zhao, L. Lei, and H. Zou, "Multi-scale object detection in remote sensing imagery with convolutional neural networks," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 145, pp. 3–22, 2018.

[3] A. Goswami, D. Sharma, H. Mathuku, S. M. P. Gangadharan, C. S. Yadav, S. K. Sahu, M. K. Pradhan, J. Singh, and H. Imran, "Change detection in remote sensing image data comparing algebraic and machine learning methods," *Electronics*, vol. 11, no. 3, p. 431, 2022.

[4] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (IEEE), pp. 779–788, 2016.

[5] S. Liu, F. Li, H. Zhang, X. Yang, X. Qi, H. Su, J. Zhu, and L. Zhang, "DAB-DETR: Dynamic anchor boxes are better queries for DETR," in *International Conference on Learning Representations*, 2022.

[6] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Computer Vision – ECCV 2020*, (ECCV), pp. 213–229, 2020.

[7] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "Centernet: Keypoint triplets for object detection," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, (IEEE), pp. 6568–6577, 2019.

[8] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[9] C. Zhang, J. Su, Y. Ju, K.-M. Lam, and Q. Wang, "Efficient inductive vision transformer for oriented object detection in remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[10] C. Zhang, K.-M. Lam, and Q. Wang, "Cof-net: A progressive coarse-to-fine framework for object detection in remote-sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–17, 2023.

[11] G. Ghiasi, T.-Y. Lin, and Q. V. Le, "Nas-fpn: Learning scalable feature pyramid architecture for object detection," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (IEEE), pp. 7029–7038, 2019.

[12] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (IEEE), pp. 936–944, 2017.

[13] G. Jocher, A. Stoken, J. Borovec, A. Chaurasia, L. Changyu, A. Hogan, J. Hajek, L. Diaconu, Y. Kwon, Y. Defretin, *et al.*, "ultralytics/yolov5: v5. 0-yolov5-p6 1280 models, aws, supervise. ly and youtube integrations," *Zenodo*, 2021.

[14] H. Dong, J. Pan, L. Xiang, Z. Hu, X. Zhang, F. Wang, and M.-H. Yang, "Multi-scale boosted dehazing network with dense feature fusion," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (IEEE), pp. 2154–2164, 2020.

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

[16] W. Liu, G. Ren, R. Yu, S. Guo, J. Zhu, and L. Zhang, "Image-adaptive yolo for object detection in adverse weather conditions," in *AAAI Conference on Artificial Intelligence*, vol. 36, pp. 1792–1800, 2022.

[17] C. Yang, M. Jin, X. Jia, Y. Xu, and Y. Chen, "Adaint: Learning adaptive intervals for 3d lookup tables on real-time image enhancement," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17501–17510, 2022.

[18] B. Li, W. Ren, D. Fu, D. Tao, D. Feng, W. Zeng, and Z. Wang, "Benchmarking single-image dehazing and beyond," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 492–505, 2018.

[19] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 159, pp. 296–307, 2020.

[20] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, (IEEE), pp. 1956–1963, 2009.

[21] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *2017 IEEE International Conference on Computer Vision (ICCV)*, (IEEE), pp. 2999–3007, 2017.

[22] M. Hnewa and H. Radha, "Multiscale domain adaptive yolo for cross-domain object detection," in *2021 IEEE International Conference on Image Processing (ICIP)*, (IEEE), pp. 3323–3327, 2021.

[23] B. Li, X. Peng, Z. Wang, J. Xu, and D. Feng, "Aod-net: All-in-one dehazing network," in *Proceedings of the IEEE international conference on computer vision*, pp. 4770–4778, 2017.

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

[25] X. Qin, Z. Wang, Y. Bai, X. Xie, and H. Jia, "Ffa-net: Feature fusion attention network for single image dehazing," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, pp. 11908–11915, 2020.

[26] B. Li, W. Ren, D. Fu, D. Tao, D. Feng, W. Zeng, and Z. Wang, "Benchmarking single-image dehazing and beyond," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 492–505, 2018.

[27] D. Zhao, B. Mo, X. Zhu, J. Zhao, H. Zhang, Y. Tao, and C. Zhao, "Dynamic multi-attention dehazing network with adaptive feature fusion," *Electronics*, vol. 12, no. 3, p. 529, 2023.

[28] Y. Wang, X. Yan, K. Zhang, L. Gong, H. Xie, F. L. Wang, and M. Wei, "TogetherNet: Bridging Image Restoration and Object Detection Together via Dynamic Enhancement Learning," *Computer Graphics Forum*, 2022.