drones

MDPI

Article

WRRT-DETR: Weather-Robust RT-DETR for Drone-View Object Detection in Adverse Weather

Bei Liu †, Jiangliang Jin *,†, Yihong Zhang * and Chen Sun

College of Information Science and Technology, Engineering Research Center of Digitalized Textile and Fashion Technology, Ministry of Education, Donghua University, Shanghai 201620, China;
2232296@mail.dhu.edu.cn (B.L.); 2232203@mail.dhu.edu.cn (C.S.)
* Correspondence: jinjiangliang@dhu.edu.cn (J.J.); zhangyh@dhu.edu.cn (Y.Z.)
† These authors contributed equally to this work.

Abstract: With the rapid advancement of UAV technology, robust object detection under adverse weather conditions has become critical for enhancing UAVs' environmental perception. However, object detection in such challenging conditions remains a significant hurdle, and standardized evaluation benchmarks are still lacking. To bridge this gap, we introduce the Adverse Weather Object Detection (AWOD) dataset—a large-scale dataset tailored for object detection in complex maritime environments. The AWOD dataset comprises 20,000 images captured under three representative adverse weather conditions: foggy, flare, and low-light. To address the challenges of scale variation and visual degradation introduced by harsh weather, we propose WRRT-DETR, a weather-robust object detection framework optimized for small objects. Within this framework, we design a gated single-head global–local attention backbone block (GLCE) to fuse local convolutional features with global attention, enhancing small object distinguishability. Additionally, a Frequency–Spatial Feature Augmentation Module (FSAE) is introduced to incorporate frequency-domain information for improved robustness, while an Attention-based Cross-Fusion Module (ACFM) facilitates the integration of multi-scale features. Experimental results demonstrate that WRRT-DETR outperforms SOTA methods on the AWOD dataset, exhibiting superior robustness and detection accuracy in complex weather conditions.

Keywords: adverse weather conditions; object detection; AWOD dataset; cross fusion of attention guiding; RT-DETR

Academic Editor: Pablo Rodríguez-Gonzálvez

Received: 18 April 2025
Revised: 12 May 2025
Accepted: 13 May 2025
Published: 14 May 2025

Citation: Liu, B.; Jin, J.; Zhang, Y.; Sun, C. WRRT-DETR: Weather-Robust RT-DETR for Drone-View Object Detection in Adverse Weather. *Drones* **2025**, *9*, 369. https://doi.org/10.3390/drones9050369

1. Introduction

Recent advancements in deep learning and large-scale datasets have significantly improved object detection from UAV perspectives [1]. However, real-world adverse weather conditions, rain, snow, low light, fog, etc., continue to pose substantial challenges by degrading visual quality and impairing the reliability of computer vision systems [2,3]. These limitations are particularly pronounced in maritime scenarios, including civilian surveillance, environmental monitoring, military reconnaissance, and search-and-rescue operations. Enhancing detection robustness under such conditions has therefore emerged as a critical research priority.

As illustrated in Figure 1, drone-based object detection is hampered by two fundamental challenges. First, object appearance diversity arises from significant scale variations due to changes in UAV altitude and viewing angle, alongside large intra-class discrepancies (e.g., frontal vs. side views). Additionally, small, ambiguous targets captured from high altitudes exacerbate detection difficulty. Second, environmental degradations—such as

fog, low-light conditions, and flare, severely obscure discriminative features, undermining reliable recognition.

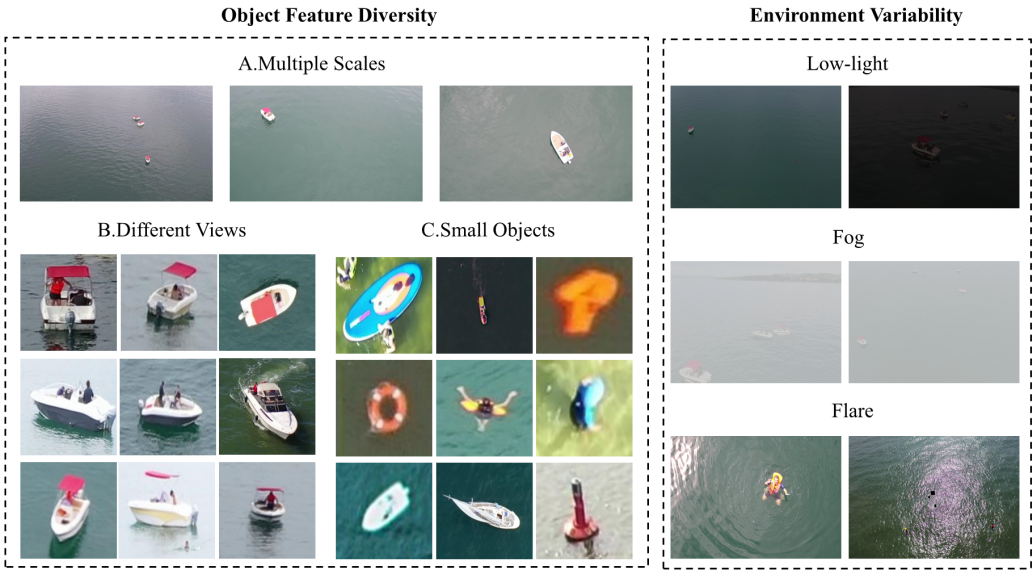


Figure 1. Challenges of drone object detection in adverse weather.

Despite the availability of datasets like SeaDronesSee [4] and MOBDrone [5], which are primarily focused on clear-weather conditions, collecting real-world UAV data under diverse weather conditions is both costly and operationally difficult. Moreover, unpredictable maritime dynamics, such as wave interference, fog particle scattering, and occlusion, further complicate the task of robust object detection [6].

To mitigate data scarcity, recent advancements in synthetic weather generation provide a viable solution [7,8]. However, many existing methods for synthetic weather generation rely on simplistic techniques, which fail to accurately reflect the physical properties of real weather conditions. To overcome this limitation, we leveraged monocular depth estimation to generate depth information, which synthesizes images with realistic weather degradation, supplementing the lack of real-world data. In addition, we introduce a novel metric, the Weather Degradation Rate (WDR), which combines the degree of image distortion and structural similarity [9] to quantitatively evaluate the degree of image degradation.

Although image restoration techniques [10,11] offer a potential pathway for mitigating visual degradation, they often introduce artifacts or over-smoothing effects, which are particularly harmful when detecting small, occluded maritime targets. These limitations indicate that robust detection under complex weather cannot rely on image enhancement alone.

To address these challenges, we introduce the AWOD dataset, a large-scale UAV maritime benchmark specifically designed for adverse weather conditions, and propose a novel weather-resilient detection framework, WRRT-DETR. In summary, our key contributions are follows:

- Construction of the AWOD: The AWOD dataset is a large-scale dataset constructed from a UAV perspective aimed at solving object detection difficulties in adverse weather.
- The Gated global–local attention backbone network: The GCLE block integrates depthwise convolution, pooled transposed attention, and gated attention to efficiently fuse global and local information, enhancing object perception while reducing computational complexity and improving model robustness and detection accuracy in complex environments.

- Spatial–Frequency Augmented Enhancement (FSAE) module: By integrating frequency and spatial domain information, global frequency features compensate for the missing local spatial information, thereby strengthening the model's capacity to detect occluded and low-contrast objects in complex environments.
- Attention-Guided Cross-Fusion Module (ACFM): This is designed to aggregate features from different stages while assigning importance weights to them. This module effectively filters out redundant information and background interference, enhancing the model's ability to represent object features in complex environments.

2. Related Work

2.1. Drone-View Datasets

In recent years, with the widespread application of drone technology, there has been a rapid advancement in research on object detection from drone perspectives. To support this, several dedicated datasets have been proposed. One such dataset, BirdsEyeView [12], includes 70 videos and 5000 static images captured from various sources, covering different scenes, perspectives, and altitudes, reflecting a wide range of real-life scenarios. The TinyPerson [13] dataset focuses on beach and sea scenes, with a particular emphasis on detecting pedestrians by the sea. SeaDronesSee [4] is one of the most widely used datasets for detecting and tracking maritime objects, covering primarily sea-based objects. VisDrone2019 [14] is one of the most widely used UAV-based datasets, comprising 10,209 static images that cover a variety of scenes, weather conditions, and lighting environments. It includes diverse object categories such as pedestrians, cars, bicycles, and tricycles. However, most existing datasets, including VisDrone2019, are primarily focused on urban and traffic environments, with limited coverage of imagery captured by maritime UAVs. As drone applications increasingly extend into complex and harsh environments, there is an urgent need for datasets that reflect detection in such complex conditions. To meet this demand, the RTTS [15] dataset was introduced, which includes real-world hazy images collected from traffic surveillance systems, representing urban traffic scenes under various weather conditions. The BDD100K [16] dataset is a large-scale dataset that covers images under various weather and time conditions, primarily focusing on complex urban environments and foggy cityscapes [17]. The image is generated by simulating a foggy environment with a cityscape dataset. However, these datasets do not address the object detection needs in drone-captured scenes. To solve this issue, the AWOD dataset was introduced, which contains images under diverse atmospheric and illumination scenarios to mirror the multifaceted and intricate circumstances in actual settings, thereby addressing a crucial missing part in drone-view object detection.

2.2. Adverse Weather Object Detection

In complex environments, real-time vehicle detection based on drones has garnered significant attention. Wang et al. proposed a deep learning framework optimized for aerial traffic monitoring [18]. Li and Xu enhanced the robustness against visual degradation by employing dataset augmentation and multi-scale detection strategies [19]. However, when it comes to the task of drone detection in complex maritime environments, detectors trained on clean images often fail to deliver satisfactory performance under adverse weather conditions (e.g., rainy, foggy, or low-light environments). This is due to the domain shift between the training and testing data [20]. To address this challenge, three primary approaches have been proposed. The first approach mitigates the impact of adverse weather on images through pre-processing, such as deraining [21], desnowing [22], dehazing [23,24] or low-light enhancement [25,26]. These methods typically rely on strong pixel-level supervision and can achieve promising results. However, they often remove potential

fine details and image textures, along with weather-related information, which limits their detection performance in real-world scenarios. The second approach involves joint restoration and detection [27]. Dual-branch networks are utilized to perform both image restoration and object detection concurrently, with the two branches sharing a feature extraction module. While this approach can enhance detection performance by improving image quality, balancing the weights of the restoration and detection tasks during training presents a significant challenge, often leading to imbalanced performance between the two tasks. The third approach employs unsupervised domain adaptation, which reduces domain discrepancies by aligning the features of clean images (source domain) and adverse weather images (the object domain) [20]. This technique eliminates the need for extensive labeled data in the object domain. However, it may overlook critical latent information necessary for enhancing detection effectiveness throughout the image restoration process.

2.3. Drone-View Object Detection

Due to the varying flight heights and angles of drones, there is a significant difference in object scales within drone-captured images, often leading to the presence of small objects. Multi-scale feature fusion and small-object enhancement techniques are regarded as critical approaches for improving small-object detection capabilities and overall detection accuracy [28]. By integrating features across different scales, multi-scale feature fusion effectively captures the fine details of small objects, while small-object enhancement techniques further optimize detection performance through targeted improvements [29].

Existing feature fusion methods include FPN [30], PANet [31], and BiFPN [32]. FPN integrates multi-scale features from different stages using a top-down unidirectional pathway. PANet builds upon FPN by employing a multi-directional fusion strategy, further enhancing feature integration through the addition of a bottom-up pathway. BiFPN introduces a bidirectional cross-scale feature fusion approach, enabling efficient and lightweight multi-scale feature aggregation. Additionally, DFLFFN utilizes deep feature learning and feature fusion networks, improving small object detection [33], while Zhu et al. use global multi-level perception and dynamic region aggregation [34]. YOLOv7-sea enhances the detection capability for small maritime targets by incorporating an additional prediction head and the SimAM attention module [35], whereas YOLOv8n-Tiny improves the model's ability to detect small objects at sea by refining the YOLOv8 architecture [36].

3. AWOD Dataset

3.1. Dataset Introduction

Benchmark datasets have long been crucial for evaluating and comparing the performance of various object detection methods. However, the scarcity of publicly available datasets for maritime object detection imposes significant limitations on advancing research in this area. To narrow this gap, we constructed a new adverse weather maritime object detection dataset, referred to as the Adverse Weather Object Detection (AWOD) dataset.

To enrich the AWOD dataset, we selected visually clean and diverse images from the publicly available SeaDronesSee [4] and MOBDrone [5] datasets. These images were chosen not to optimize detection performance, but to ensure compatibility with weather synthesis techniques. Specifically, we avoided images with pre-existing environmental interference (e.g., strong light or noise), as these could compromise the realism of synthetic degradation, resulting in a subset of 10,000 normal-weather drone images, named PureSea.

Based on the PureSea dataset, we synthesized three typical adverse weather conditions—foggy, low light, and flare—using a self-supervised monocular depth estimation method. Specifically, 6750 images were randomly selected for both foggy and low-light

scenarios, while 6500 images were chosen for flare conditions. These synthesized weather-degraded images were then combined to form the proposed AWOD dataset.

The established AWOD dataset, as shown in Figure 2, contains uniformly distributed images degraded by three weather conditions: foggy, low light, and flare (light reflection), totaling 20,000 images. To facilitate public research, AWOD is annotated with six common class labels: ignore , swimmer, boat, jetski, life_saving_appliances, buoy.

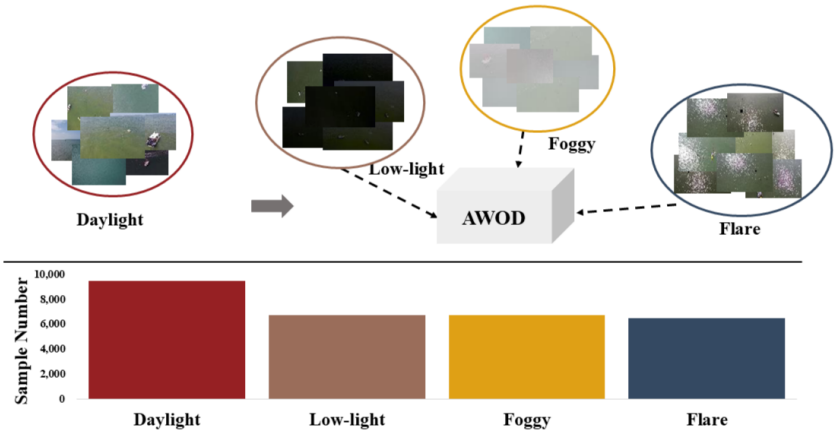


Figure 2. Overview of the established AWOD benchmarks. The dataset includes three adverse weather conditions (foggy, low light, and flare); these degraded images are synthesized from clean pictures, providing 20,000 degraded images in total.

As shown in Table 1, we counted the number of large, medium, and small objects in these five types of labels, following the classification method proposed by the COCO dataset. In COCO, a small object is defined as a box with an area of fewer than 32 $\times$ 32 pixels; medium with between 32 $\times$ 32 and 96 $\times$ 96 pixels; and large with greater than 96 $\times$ 96 pixels.

Table 1. AWOD dataset statistics.

Class	Instances	Object Size		
		Small	Medium	Large
Swimmer	83,037	82,445	376	216
Boat	29,156	24,663	2495	1998
Buoy	9731	9689	42	0
Jetski	5219	4899	253	67
Life_saving_appliances	2068	2068	0	0

3.2. Synthetic Weather Degradation

In real-world applications, images captured by camera systems are often affected by complex weather conditions such as fog, flares, and snow [37,38]. These factors not only significantly degrade image quality but also pose challenges to subsequent image processing and visual analysis tasks. Although rain and snow are common weather phenomena, in practical UAV inspection applications, flight missions are generally not carried out under heavy rain or snow conditions due to safety and visibility concerns. Therefore, in constructing the maritime weather dataset, we focus primarily on simulating conditions such as fog, low light, and flares while excluding rain and snow from the simulation scope. As shown in Table 2, compared to previous marine remote sensing datasets, AWOD encompasses data from common maritime weather conditions. Figure 3 illustrates images generated via three synthesis methods across 12 typical environments.

Table 2. A comparison between our AWOD dataset and commonly used drone-view datasets.

Dataset	Object Class	Images	Adverse Weather
BirdsEyeView (2019) [12]	6	5k	-
TinyPerson (2019) [13]	2	2k	-
SeaDronesSee (2022) [4]	6	10k	-
VisDrone (2022) [39]	10	10k	-
ShipDataset (2023) [40]	1	18k	-
AWOD (ours)	6	20k	fog, low light, flares

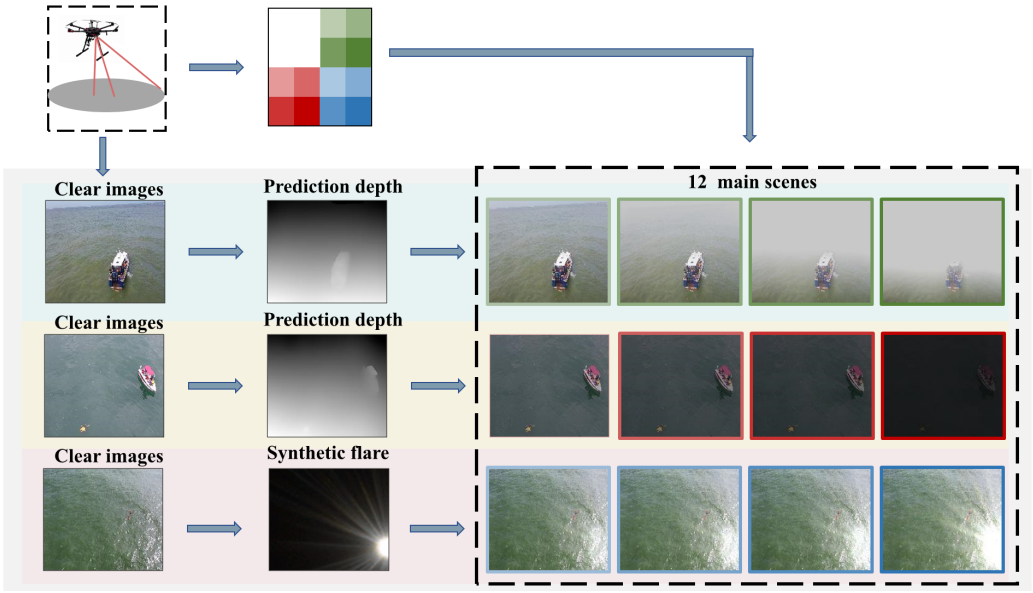


Figure 3. Visual diversity of synthetic weather. Three weather simulation methods and twelve representative types in the AWOD dataset. Fog leads to partial or full occlusion of distant objects, reducing overall visibility. Low light conditions diminish contrast and obscure fine-grained details, challenging object recognition. Flare introduces localized overexposure, typically around reflective surfaces, which distorts image information.

Based on the atmospheric scattering model [41] and the retinal vision theory [42], we define the clear image $I(x, y)$, the depth map $D(x, y)$, the atmospheric light A and the transmission map T. Our image synthesis process is formulated as follows:

$$T(x, y) = e^{-\beta D(x,y)} \tag{1}$$

$$I_{\text{low-light}}(x, y) = I(x, y) + \alpha \cdot D(x, y) \tag{2}$$

$$I_{\text{foggy}}(x, y) = T(x, y) \cdot I(x, y) + A \cdot (1 - T(x, y)) \tag{3}$$

Among these, $I_{\text{foggy}}(x, y)$ and $I_{\text{low-light}}(x, y)$ map the clear image to images affected by fog and low light. Since the PureSea maritime dataset lacks available depth maps, we utilized the pre-trained depth estimation model DPT [43] to predict depth information $D(x, y)$. The parameter β denotes the attenuation coefficient controlling the foggy density, which is set to 1.0. A represents the atmospheric light, across the value range [0.3, 0.7] [41]; however, a true foggy image has a noticeable flare effect, so in our formula, the ambient light A changes depending on the local brightness of the clear image $I(x, y)$, while α is a function of adjusting the intensity of the light based on depth.

For flare image generation, leveraging the additive nature of light, we first constructed scattered flare images and superimposed them onto the original images to simulate the

degradation effect caused by flare. A detailed description of the degradation method is provided below.

When sunlight reflects off calm seawater at specific angles, the sea surface acts as a mirror, reflecting light [44]. This reflection, combined with unintended scattering within the camera lens, results in flare. Flare primarily consists of bright spots (flare) and streaks. To generate scattered flare images, this study models and synthesizes these two components separately. Flare spots exhibit a high-brightness center surrounded by a smooth gradient halo. We adjusted gradient patterns to synthesize realistic flare spots. Streaks, often caused by scratches or oil smudges on the lens, were simulated using Gaussian functions. A feathering template was applied to adjust the brightness around the streaks.

Finally, the synthesized flare spots and streaks were combined into complete scattering flare images, which were overlaid onto the original images to produce degraded effects. Random noise and gain were added to enhance realism and cover a wide range of noise levels observed in real-world scenarios:

$$I_{\text{flare}}(x, y) = I(x, y) + F + G(0, \sigma^2) \tag{4}$$

The clear images $I(x, y)$ were sampled from the PureSea dataset, processed with gamma correction, basic image transformations, random noise, and gain addition, while constraining pixel value ranges. Flare images F were extracted from the synthetic dataset and enhanced with gamma correction, background removal, color jittering, and blurring.

3.3. Weather Degradation Ratio

According to the lack of evaluation benchmarks, quantitatively assessing the impact of weather-induced degradation is challenging even for humans. We hypothesize that the Weather Degradation Rate (WDR) depends on the degree of distortion in the image and the occlusion of objects caused by the weather. Suppose the two images to be compared are the original image x and the degraded image y (size M×N); therefore, WDR can be expressed as follows:

$$S_1(x, y) = \frac{2\sigma_{xy} + c_1}{\sigma_x^2 + \sigma_y^2 + c_2} \tag{5}$$

where μ_x and μ_y represent the mean luminance of images x and y, respectively; the terms σ_x^2 and σ_y^2 represent their corresponding variances; while σ_{xy} indicates their covariance. A small constant C_1 is introduced to prevent division by zero, and Equation (5) is used to calculate image contrast and structural differences.

$$S_2(x, y) = 1 - \frac{\|\mu_x - \mu_y\| + \|\sigma_x - \sigma_y\|}{\|\mu_x\| + \|\sigma_y\|} \tag{6}$$

Equation (6) calculation measures the difference in brightness and contrast between the original image and the degraded image. Here, $\|\cdot\|$ denotes the Euclidean norm (i.e., the L_2 norm) of a vector.

$$WDR = 100(w_1 s_1 + (1 - w_1)s_2) \tag{7}$$

Ultimately, Equation (7) is obtained by adding the two parts, and w_1 represents the weight of contrast and structural differences.

To achieve fair grading, we introduce an index, the WDR, to quantitatively evaluate the degree of degradation in the AWOD images. The images are classified into the following

difficulty levels: 0∼25 to a particularly difficult level, 26∼50 to a difficult level, 51∼75 to a normal level, and 76∼100 to an easy level.

4. Materials and Methods

In recent years, end-to-end detectors have made significant progress and have been widely applied in outdoor vision systems, such as surveillance and object detection. However, their performance is severely impacted under adverse weather conditions, often failing to produce satisfactory results. Various detectors trained on common weather datasets drop significantly when evaluated on datasets affected by environmental factors (e.g., rain, flare, and low-light scenarios) [11,45].

Aside from the performance degradation caused by severe image degradation, we observed that maritime drone images often contain a high proportion of small objects. These small objects are more likely to be occluded or partially covered in scenarios with drastic lighting changes, complex backgrounds, or numerous distractors, making them easier to miss during detection [46,47] . Based on this observation, we enhanced RT-DETR by incorporating a gated single-head global–local attention backbone block, a GLCE block, a feature enhancement module, FSAE, and an attention-guided feature fusion module, ACFM; the network structure is shown in Figure 4.

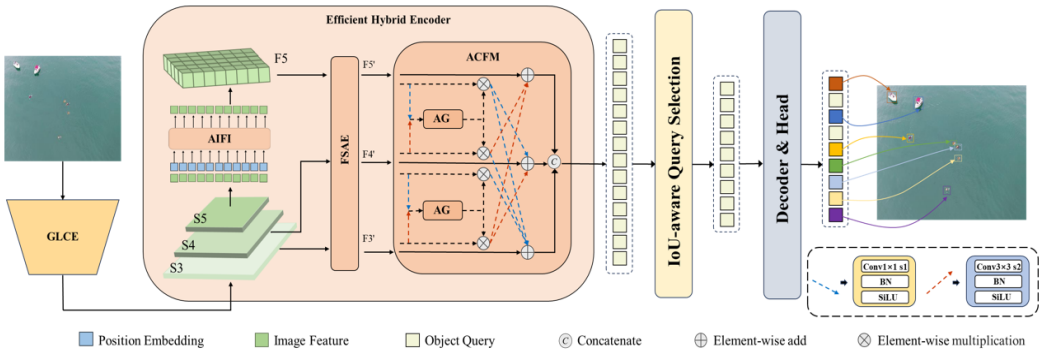


Figure 4. WRRT-DETR main network structure.

4.1. The Gated Global–Local Attention Backbone Network

ResNet-18 [48], as the backbone network, efficiently extracts features with a low parameter count and computational requirements through residual connections, making it well suited for small object detection in drone applications. In harsh environments, objects often suffer from occlusion and are difficult to distinguish from the background, especially for small objects. To enhance the recognizability of small objects in complex backgrounds, we designed the GCLE block. This block integrates gating mechanisms, local convolutions, and global attention to effectively combine local details with global context. By optimizing computational redundancy, GCLE significantly improves the performance of small object detection in adverse weather. The GCLE block structure is shown in Figure 5.

To effectively capture local and global contextual information, we divide the channels into S_{i-l} and S_{i-g} and apply different operations to each. Specifically, we perform a 1×1 convolution on the input feature S_{i-g} and decompose it into four components: z_i, Q_i, K_i, and V_i. Here, Q_i, K_i, and V_i serve as the attention mechanism, while Z_i is responsible for multi-channel feature aggregation. To improve the robustness of the features and reduce the sensitivity to minor variations, we apply Max pooling and Avg pooling to Q_i and K_i, respectively, allowing for more effective extraction of critical information.

$$X_{i-l}, X_{i-g} = \text{Split}(X_i) \tag{8}$$

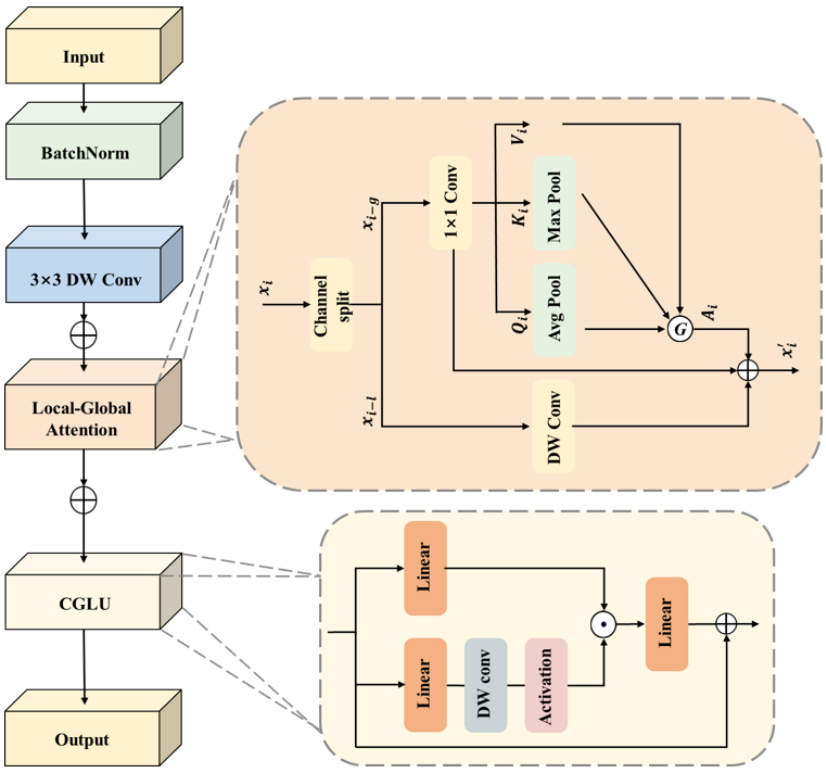


Figure 5. Global–local attention backbone network block.

Then, we apply transposed attention ($\mathcal{G}$) to embeddings Q_i, K_i, and V_i over a subset of feature channels ($\mathcal{C}/4$), ensuring linear complexity with respect to the number of tokens. The operations in our PT attention mechanism, used to derive the feature representation A_i, can be summarized in Equation (9).

$$A_i = V_i \times \left[\sigma\left(Max(K_i{}^T) \times Avg(Q_i)\right)\right] \tag{9}$$

$$x_i' = \text{concat}(DWconv(x_{i-l}), z_i, A_i(Q, K, V)) \tag{10}$$

To capture local contextual information in both spatial and channel dimensions, we apply a 3 × 3 depthwise convolution to the input feature x_{i-l}. In Equation (10), we aggregate features with varying receptive fields to aggregate both local and global contextual information. The concatenation operation is employed to merge these features, resulting in a rich global–local contextual representation.

4.2. Frequency–Spatial Augmented Enhancement

Object detection under adverse weather faces challenges such as occlusion, deformation, and environmental interference, which makes traditional spatial domain methods less effective. Frequency enhancement techniques leverage global frequency information to compensate for missing spatial details, improving robustness in detecting occluded and low-contrast objects [49].

To address this, we propose the Frequency–Spatial Augmented Enhancement (FSAE) module, which integrates frequency-domain features, contextual information, and spatial textures to enhance object representation. By optimizing computational redundancy, FSAE strengthens fine-grained features with minimal memory overhead, making it particularly suitable for small object detection on UAVs under adverse weather conditions. The structure of the FSAE module is illustrated in the accompanying Figure 6.

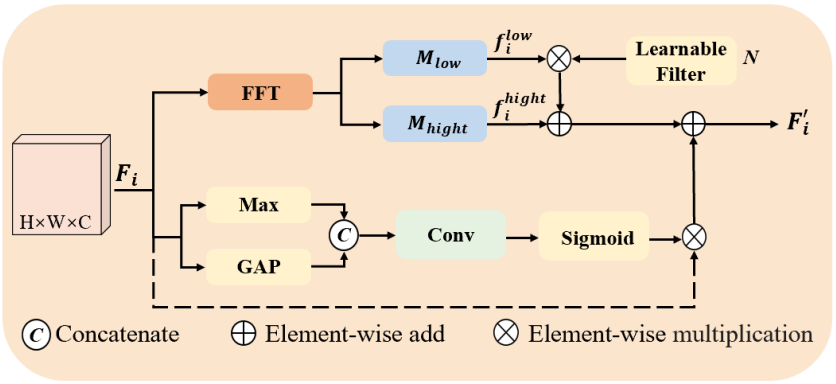


Figure 6. Frequency–Spatial Augmented Enhancement module.

Let the input feature map be denoted as $\mathcal{F} \in \mathbb{R}^{C\times H\times W}$, where C represents the number of channels and H and W correspond to the height and width, respectively. We use the Fourier transform to convert it to the frequency domain f_i, as shown in Equation (11). In the spatial domain, $F_i(x, y)$ represents the pixel values of the original features.

$$f_i(U, V) = \sum_{x=0}^{H-1}\sum_{y=0}^{W-1} F_i(x, y)e^{-j2\pi\left(\frac{Ux}{H}+\frac{Vy}{W}\right)} \tag{11}$$

We use two masks, M_{height} and M_{low}, to split f_i into a f_{low} and f_{height}. At $\mathcal{M} \in \mathbb{R}^{C\times H\times 1}$, M_{low}, the square with the side length n added is centered on the mask and assigned a value of 1, while the rest of the area is assigned a value of 0, and, conversely, the center of the mask M_{height} is 0. We then multiply between f and M to obtain the f_{low} and f_{height} components, respectively, as shown in Equation (13).

$$f_i^{low} = f_i \otimes M_{low} \tag{12}$$

$$f_i^{height} = f_i \otimes M_{height} \tag{13}$$

f_{height} values are typically associated with the details, edges, and textures of an image, while f_{low} values represent the overall structure and background. To improve feature extraction and control the influence of the background, we adaptively adjust the obtained f_{low} components with the learned filter $\mathcal{N}$, as depicted in Equation (13).

$$f_i' = f_i^{height} \oplus (f_i^{low} \otimes N) \tag{14}$$

Afterward, the adjusted low-frequency components are fused with f_{height} to produce the frequency feature map. Subsequently, the frequency feature map, following the modifications in the frequency domain, is converted back to the spatial domain via the inverse Fourier transform to derive the enhanced image representation. The transformation formula is shown in Equation (15).

$$\hat{F}_i(x, y) = \frac{1}{HW}\sum_{U=0}^{H-1}\sum_{V=0}^{W-1} f_i'(U, V)e^{j2\pi\left(\frac{Ux}{H}+\frac{Vy}{W}\right)} \tag{15}$$

We apply the spatial attention mechanism on the input feature map to further enhance feature learning. Ultimately, we fuse the feature maps generated by the frequency domain branch and the spatial domain branch to obtain the enhanced feature map, using fusion formulas such as Equation (16).

$$F_i' = \hat{F}_i \oplus (F_i \otimes \sigma\text{conv}(\text{concat}(\text{GAP}(F_i), \max(F_i)))) \tag{16}$$

4.3. Attention-Guided Cross-Fusion

In degraded images, excessive noise can obscure object features and blur boundary information. Moreover, small object features are typically less prominent, making them more susceptible to being overshadowed by background or other interference. To address this, we designed a Cross-Fusion (CF) module that effectively integrates spatial details and semantic information. The CF module accepts input feature maps from different backbone layers and distributes the fused output to multiple detection heads. By directly accessing low-level features, the CF module enhances the detection of small objects. Assume the input feature map as $\mathcal{F}$. One of the feature enhancement processes is as follows:

$$O_1 = \mathcal{F}_1 \otimes (\mathcal{F}_2 \otimes w_1(\mathcal{F}_1, \mathcal{F}_2)) \oplus (\mathcal{F}_2 \otimes w_2(\mathcal{F}_2, \mathcal{F}_3)) \tag{17}$$

where "$\oplus$" refers to element-wise summation and "$\otimes$" indicates element-wise multiplication. w_1 and w_2 represent the weights of the attention layers. Equation (17) is the upbranch of the ACFM outputs.

To better facilitate the fusion of features across different layers, we propose an attention-guided module that focuses on salient information while suppressing irrelevant background noise. This approach effectively captures small objects in complex scenes. Recognizing that global average pooling (GAP) [50] may dilute or ignore small object features in global information, we apply pooling operations along the horizontal and vertical orientations of the input features, as shown in Figure 7. This enables the finer-grained capture of small object features while reducing background interference.

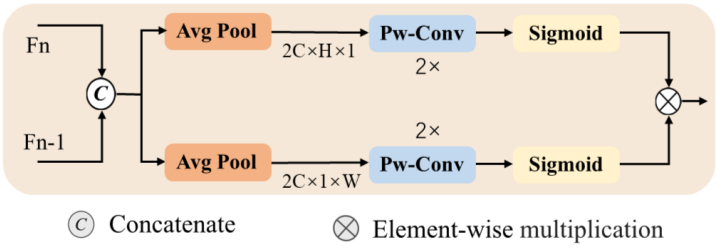


Figure 7. Channel Attention Guiding module.

The operations are defined as

$$g_u^h = \frac{1}{H} \sum_{i=1}^{H} \mathcal{F}_u(h, i) \tag{18}$$

$$g_u^w = \frac{1}{W} \sum_{j=1}^{W} \mathcal{F}_u(j, w) \tag{19}$$

which represent the input features of pooling kernels along two spatial ranges, with dimensions $H \times 1$ and $1 \times w$, respectively. These features are passed through a channel attention module composed of convolution and pooling operations to generate final fusion weights. The weight generation process is described as

$$w^h = \sigma\left(Conv\left(\delta\left(Conv\left(g^h \oplus g^w\right)\right)^h\right)\right) \tag{20}$$

$$w^w = \sigma\left(Conv\left(\delta\left(Conv\left(g^h \oplus g^w\right)\right)^w\right)\right) \tag{21}$$

Conv represents convolution layers and σ and δ represent the sigmoid function and Rule activation function, respectively. The enhanced features are merged along the channel dimension and subsequently input into parallel channel attention and spatial attention modules.

$$w = w^h \otimes w^w \tag{22}$$

Finally, to achieve more salient feature representation, the attention weights w^h and w^w are expanded and applied as scaling factors.

5. Experiments and Discussion

5.1. Evaluation Metrics

Precision (P), recall (R), and mean average precision (mAP) are used as evaluation metrics to quantitatively evaluate and compare our method. The calculation formulas are as follows:

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{23}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{24}$$

In the formulas, θ represents the Intersection over Union (IoU), TP stands for the number of predicted bounding boxes, FP is the number of predicted bounding boxes, FN is the number of unpredicted ground truth bounding boxes, and N is the number of object classes.

$$\mathrm{AP} = \int_0^1 P(R)\, dR \tag{25}$$

$$mAP_{50} = \frac{1}{N} \sum_{i=1}^{N} AP_i(IOU_{thresh} = 0.5) \tag{26}$$

$$mAP_{50:95} = \frac{1}{10} \sum_{j} \frac{1}{N} \sum_{i=1}^{N} AP_i(IOU_{thresh} = j) \tag{27}$$

The experiment uses mAP as an algorithm evaluation metric, where mAP50 represents the average detection accuracy, while mAP50:95 represents the average accuracy of all classes, with an interval of 0.05, calculated with an IoU threshold ranging from 0.5 to 0.95.

Params and GFLOPS measure model size and computational cost, respectively. Lower values favor deployment on resource-limited platforms such as UAVs. Frames per second (FPS) reflects inference speed, where higher FPS supports real-time detection.

5.2. Implementation Details

To ensure experimental consistency, all experiments were conducted on a single NVIDIA RTX 4090. The input image resolution was set at 640 × 640, and the model was trained for 100 epochs. The detailed settings are shown in Table 3.

Table 3. Hardware configuration and model parameters.

Type	Version	Type	Value
GPU	RTX 4090	Optimizer	AdamW
Python	3.8.0	Batch	16
Pytorch	1.10.0	Learning rate	1×10^{-4}
CUDA	11.3	Momentum	0.9

5.3. Performance of Detectors on AWOD

To evaluate the impact of AWOD pre-training on object detection performance in real-world complex weather conditions, we designed a series of experiments comparing three pre-training strategies: (1) directly using the MS COCO pre-trained weight as a pre-trained weight 1 released by the official RT-DETR; (2) based on pre-trained weight 1, training on the AWOD dataset for 50 epochs to obtain pre-trained weight 2; and (3) based on pre-trained weight 1, training on the AWOD dataset for 100 epochs to obtain the pre-trained weight 3. Subsequently, based on these three pre-trained weights, we trained for 100 epochs on three publicly available datasets of real weather scenarios—RTTS [15], BDD100K [16], and VisDrone2019 [14]—using mAP50 and mAP50:95 as evaluation metrics. The experimental results are shown in Table 4.

Table 4. Experimental results before and after AWOD training.

Epochs	Method	RTTS [15]		BDD-100K [16]		VisDrone2019 [14]	
		mAP50	mAP50:95	mAP50	mAP50:95	mAP50	mAP50:95
0	RT-DETR	64.0	36.6	60.1	32.6	45.8	27.7
	WRRT-DETR	66.4	37.6	62.7	33.9	49.5	29.2
50	RT-DETR	65.6	37.1	60.8	32.9	47.9	28.4
	WRRT-DETR	67.1	38.0	63.3	34.3	52.4	31.8
100	RT-DETR	66.1	37.5	61.7	32.4	49.7	29.4
	WRRT-DETR	67.5	38.3	63.8	34.7	54.8	32.4

The epochs in the table represent the number of rounds that were pre-trained on the AWOD dataset; the mAP50 and the mAP50:95 in the table are represented by percentage %.

Table 4 presents the trends in mAP50 and mAP50:95 scores for both RT-DETR and WRRT-DETR, which consistently increase as the number of AWOD pretraining epochs rises from 0 to 100. On the RTTS dataset, WRRT-DETR consistently outperforms RT-DETR across all pretraining stages. In particular, after 100 epochs of AWOD pretraining, WRRT-DETR reaches an mAP50 of 67.5% and an mAP50:95 of 38.3%.

On the BDD-100K autonomous driving dataset, the improvement in detection accuracy is relatively moderate due to the difference in perspective between the dataset (collected from vehicle-mounted cameras) and AWOD. Nevertheless, WRRT-DETR still exhibits steady improvements in mAP50 and mAP50:95, with gains ranging from 0.3 to 0.9. On the VisDrone2019 dataset, which shares an aerial perspective and high weather diversity with AWOD, the performance gains from AWOD pretraining are the most pronounced. WRRT-DETR's mAP50 increases from 49.5% to 54.8%, and its mAP50:95 rises from 30.2% to 32.4%. On the VisDrone2019 dataset, after 50 epochs of pretraining, RT-DETR's mAP50 improves from 45.8% to 47.9%, while WRRT-DETR's mAP50 increases from 49.5% to 52.4%. When the number of pretraining epochs increases to 100, WRRT-DETR achieves mAP50 and mAP50:95 scores of 54.8% and 32.4%, respectively, while the corresponding metrics for RT-DETR are 49.7% and 29.4%. WRRT-DETR consistently outperforms RT-DETR across all datasets and pretraining settings, demonstrating superior robustness and detection capability.

Overall, the experimental results clearly show that AWOD pretraining significantly improves object detection performance under complex weather conditions, with stronger effects observed at higher pretraining epochs. WRRT-DETR, with its greater robustness and adaptability, demonstrates superior performance in such tasks.

5.4. Ablation Experiments

Ablation experiments on the AWOD dataset were conducted to evaluate the effectiveness of the proposed method. Since model parameters and computational complexity are closely related to factors such as input image resolution, batch size, and network depth, ablation experiments were carried out under controlled conditions with consistent resolution and batch size.

We selected RT-DETR as the baseline for the ablation study, although it is not the only possible choice. To ensure fairness, all experiments used identical datasets and parameter settings. The performance was evaluated using the mean average precision mAP50 and mAP50:95 as metrics.

We progressively introduced improvements in ablation experiments to assess the impact of each module, starting with a baseline model, by adding one module at a time. Each module was quantitatively evaluated to measure its effect on detection accuracy and computational complexity. Table 5 shows that the GCLE module integrates global context information, spatial local attention, and gated attention mechanisms, effectively enhancing model performance and improving detection accuracy by 0.6%. This result underscores the crucial role of global context modeling in improving detection performance. The FSAE module significantly enhances performance in complex scenes and small object detection tasks by introducing a frequency–space-enhancement mechanism, leading to a 2.1% improvement in detection accuracy. This improvement further validates the importance of frequency features in small objects and complex environments. The ACFM module enhances model performance through cross-scale feature fusion, yielding an additional 2.7% increase in accuracy. This result indicates that effectively integrating model robustness and noise suppression capability can be improved by shallow detail features and deep semantic features. Therefore, we conclude that all of the proposed modules contribute significantly to the RT-DETR baseline detector.

Table 5. Ablation studies of the model.

GCLE	FSAE	ACFM	P (%)	R (%)	mAP50 (%)	mAP50:95 (%)
			84.9	76.8	76.9	42.3
✓			85.4	77.1	77.5	42.9
	✓		86.4	77.8	78.4	64.3
		✓	86.3	78.3	80.1	47.5
✓	✓		87.8	77.8	79.6	47.7
✓	✓	✓	89.2	80.0	82.3	48.7

5.5. Comparative Experiments

We selected YOLOv9m, YOLOv10m, YOLOv11m, DINO, DAB-DETR, DN-DETR, RT-DETR, and RTDETR-R50 for fair comparison. To validate the effectiveness of WRRT-DETR across four difficulty levels in the AWOD dataset, we conducted evaluations on each level as well as the entire dataset. To emphasize the performance of our proposed model, its scores for each metric are highlighted in bold in Table 6.

As shown in Table 6, WRRT-DETR demonstrated stable and significant advantages over SOTA methods at all difficulty levels and across the full dataset. Notably, the results reveal that the model maintains robust detection performance under mild adverse weather conditions. However, as weather conditions deteriorate, many models experience substantial performance degradation. In contrast, our model consistently achieves outstanding performance across different weather conditions, even in drone-view scenarios characterized by complex backgrounds, dense objects, and variable target sizes. This fur-

ther validates the proposed weather degradation metric's ability to accurately categorize weather-degraded images into distinct levels.

Table 6. Comparison with existing methods on AWOD. AWOD, with its four difficulty levels, provides a quantitative method to analyze the performance degradation of detectors in adverse weather conditions.

Method	Easy (6500)		Normal (3100)		Difficult (8400)		Particularly (2000)		All (20,000)	
	mAP50	mAP50:95	mAP50	mAP50:95	mAP50	mAP50:95	mAP50	mAP50:95	mAP50	mAP50:95
YOLOv9m [51]	72.4	48.3	67.8	40.5	65.1	37.0	50.7	31.3	65.4	34.1
YOLOv10m [52]	72.7	49.2	67.3	40.8	65.9	37.2	51.3	32.9	65.6	35.3
YOLOv11m [53]	73.1	50.6	68.4	41.4	66.6	37.5	51.8	33.8	66.9	36.1
DINO [54]	84.8	56.5	77.9	46.2	76.6	44.5	74.4	41.6	78.9	44.7
DAB-DETR [55]	85.4	54.3	76.1	43.9	73.1	43.0	71.3	39.4	76.3	40.7
DN-DETR [56]	84.9	54.8	76.7	44.3	73.2	42.8	70.8	39.1	74.9	39.9
RT-DETR [48]	85.5	55.8	77.4	45.8	75.8	44.1	74.2	41.6	76.9	43.3
RT-DETR-R50 [48]	86.1	56.9	78.8	46.8	76.3	44.5	74.7	41.8	77.7	45.8
WRRT-DETR (ours)	**86.3**	**56.7**	**80.2**	**47.5**	**78.8**	**45.9**	**76.4**	**43.4**	**82.3**	**46.6**

The mAP50 and the mAP50:95 in the table are represented by percentages (%).

Moreover, as shown in Table 6, the detectors suffered significant degradation across different difficulty levels. This underscores the effectiveness of our proposed gradual grading strategy, which accurately categorizes degraded images into varying levels of difficulty, facilitating deeper insights into the impact of adverse weather on object detection performance.

The results indicate that algorithms such as YOLO and DETR experience varying degrees of performance degradation in object detection as conditions and weather scenarios worsen. For most detectors, performance on lightly degraded datasets (simple scenes and low degradation levels) surpasses that on heavily degraded datasets (complex scenes and occluded objects). Despite using YOLO models with comparable parameter sizes to DETR models, their performance lags significantly behind DETR-based methods. Notably, the DINO detector exhibits superior performance on heavily degraded images due to its robust adversarial learning capabilities.

To ensure a fair comparison of state-of-the-art object detection methods, all experiments were conducted for 100 training cycles on the AWOD dataset. In addition to the YOLO and DETR series, we compared several open-source methods, including TOOD [57], YOLO-OW [58], and UAV-YOLO [59]. In Table 7, the best-performing values for each evaluation metric are shown in bold.

Table 7. Experimental results of different methods training on AWOD.

Method	Params (M)	FLOPs (G)	P (%)	R (%)	mAP50 (%)	mAP50:95 (%)	FPS (s)
TOOD [57]	32.0	199.0	55.3	49.4	58.7	34.6	34.9
YOLOv9m [51]	25.3	102.3	69.8	62.0	65.4	34.1	100.1
YOLOv10m [52]	24.3	120.0	73.9	63.4	65.6	35.3	132.5
YOLOv11m [53]	25.1	67.7	76.2	63.5	66.9	36.1	**143.7**
YOLO-OW [58]	42.1	94.8	78.1	69.8	70.5	34.9	61.3
UAV-YOLO [59]	47.4	103.3	70.9	63.1	62.7	35.7	80.9
RT-DETR [48]	**20.0**	**57.3**	83.2	76.1	76.9	43.3	71.5
RT-DETR-R50 [48]	42.1	129.9	84.9	76.8	77.7	45.8	53.5
WRRT-DETR (ours)	20.2	58.6	**86.7**	**79.5**	**82.3**	**46.6**	66.4

As shown in Table 7, we observe that YOLOv11m achieves the best performance among YOLO-based methods, with an mAP50 of 66.9% and an mAP50:95 of 36.1%. Among the two detectors specifically designed for UAV aerial imagery, YOLO-OW demon-

strates relatively better performance, achieving an mAP50 of 70.5% and an mAP50:95 of 34.9%. UAV-YOLO, on the other hand, exhibits average performance on the AWOD dataset. Notably, YOLO-OW was designed to address false detections caused by water surface flare, while UAV-YOLO was improved based on the Visdrone2019 dataset. This explains why UAV-YOLO only achieves normal performance in maritime object detection. The results indicate that WRRT-DETR outperforms other models across all major evaluation metrics. Specifically, it achieves an accuracy of 87.7% and a recall rate of 79.5%, exceeding the second-best model by 1.8% and 2.7%, respectively. These findings suggest that by integrating global and local attention mechanisms, feature enhancement in both frequency and spatial domains, and cross-scale feature fusion, WRRT-DETR enhances its capability to detect small objects under adverse weather conditions. Moreover, WRRT-DETR achieves an mAP50 and mAP50:95 of 82.3% and 46.6%, respectively, significantly surpassing other baseline models.

Among the compared models, the number of parameters (Params) is generally within the same order of magnitude, ranging from 20 M to 50 M, while the floating-point operations (FLOPs) are mostly concentrated around 100 G. On GPUs with sufficient computational resources, most object detectors achieve processing speeds exceeding 50 FPS. UAV-YOLO has the largest number of parameters at 47.4 M, whereas TOOD exhibits the lowest inference speed at only 56.3 FPS. WRRT-DETR achieves a favorable balance between computational efficiency and resource consumption. It contains only 20.2 M parameters—just 0.2 M more than RT-DETR-R18—yet delivers a 5.4% improvement in detection accuracy. Compared with models of similar parameter scales, such as YOLOv9m, YOLOv10m, and YOLOv11m, WRRT-DETR demonstrates superior detection performance, despite slightly lower inference speed. Specifically, it outperforms the three YOLOv models by 16.9%, 16.7%, and 15.4% in detection accuracy, respectively, while maintaining a real-time capable inference speed of 66.4 FPS. Overall, WRRT-DETR significantly enhances overall detection performance and offers an accurate and efficient solution for small object detection tasks.

5.6. Robustness Analysis

To evaluate the robustness of WRRT-DETR under adverse weather conditions, we perform unified training on the full AWOD dataset, ensuring broad generalization capability under diverse environmental conditions. Subsequently, we evaluate the trained model on three representative degraded subsets of fog, low light, and flare to quantify its category-wise detection performance. The only variable in this experiment is the weather condition during the testing phase, allowing us to isolate and analyze the impact of each degradation type on model performance. This setup provides valuable insight into potential category-level vulnerabilities that may arise in real-world deployment scenarios, as shown in Table 8.

Table 8. The mAP50 results of different weather conditions for five categories.

Class	RT-DETR			WRRT-DETR		
	Foggy	**Low-Light**	**Flare**	**Foggy**	**Low-Light**	**Flare**
All	70.4	69.6	77.3	77.0 (+6.6)	74.5 (+4.9)	80.5 (+3.2)
Swimmer	72.8	54.4	73.4	78.3 (+5.5)	58.6 (+4.2)	77.1 (+3.7)
Boat	95.5	92.3	96.9	96.4 (+0.9)	94.6 (+2.3)	97.2 (+0.3)
Buoy	74.8	61.4	75.9	79.0 (+4.2)	68.4 (+7.0)	81.8 (+5.9)
Jetski	86.3	83.7	85.4	90.3 (+4.0)	84.6 (+0.9)	90.9 (+1.5)
Life_saving_appliances	35.4	50.8	45.7	40.8 (+5.4)	54.4 (+3.7)	49.8 (+4.1)

In low-light conditions, insufficient illumination makes the contours of small targets like life_saving_appliances blurry and lowers their contrast with the background, causing

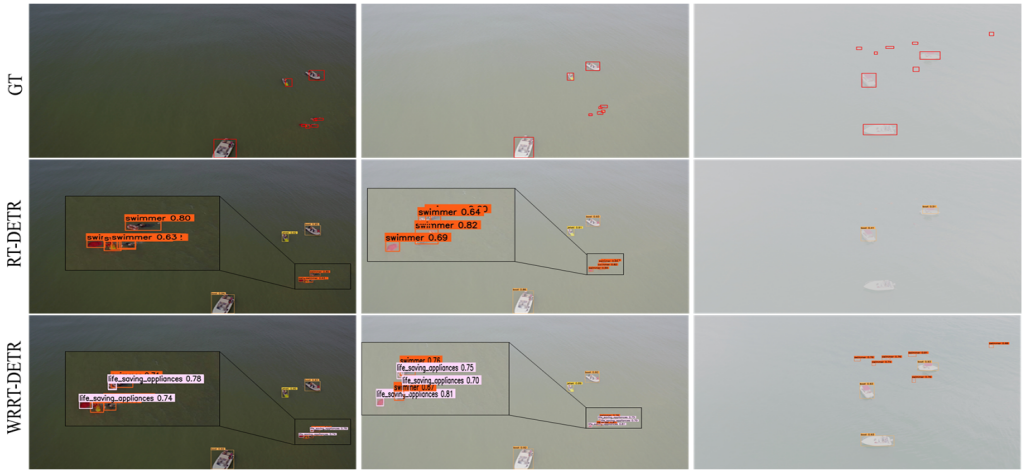
missed detection. WRRT-DETR introduces attention mechanisms, frequency domain spatial feature enhancement, and small object enhancement modules. These significantly boost the feature representation for target detection, raising mAP50 from 70.4% to 77.0%. In foggy conditions, the low contrast and occlusion make it harder to identify targets like swimmer and life_saving_appliances in overlapping areas. WRRT-DETR's improved feature fusion mechanism effectively improves the detection of occluded targets, achieving an mAP50 of 74.5%, a 4.9% increase. In flare scenes, strong light reflection and flare cause local feature degradation and artifacts, increasing false detection risks. WRRT-DETR uses cross-scale feature fusion and saliency-guided mechanisms to reduce background noise and redundant information interference. This results in a 3.2% increase in mAP50 to 80.5%, showing stable detection performance under strong interference.

Per-class analysis further reveals that detection performance is closely correlated with object size and prevalence. Large targets such as boats and jet skis maintain a high accuracy of 90% mAP50 across all weather conditions due to their distinct contours and scale. In contrast, small targets like swimmer and life_saving_appliances are more vulnerable to environmental degradation, leading to frequent missed detections under fog and low light and increased false positives under flare.

5.7. Visualization Experiments

To better illustrate the practical detection performance of the WRRT-DETR model under varying levels of degradation, we selected four images each from foggy, flare and low-light scenarios and performed inference using the trained weights. The experimental results are shown in the figures, where the rows represent the input images, the detection results from the baseline, and the detection results from WRRT-DETR, respectively.

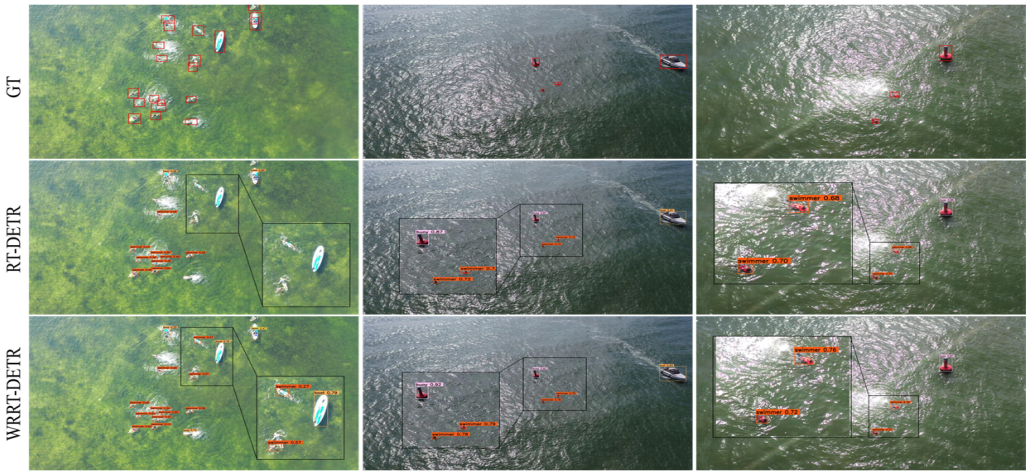
The visual analysis of the results clearly demonstrates the effectiveness of the proposed algorithm in addressing common issues in baseline methods, such as missed and false detections, as shown in Figure 8. In order to better demonstrate the missed detections and false detections, we have zoomed in on certain areas of the image. With the deepening of the degradation, the baseline model produced obvious missed detections, while our model was able to maintain stable detection performance. For instance, Figure 8c clearly shows that the baseline algorithm misclassifies "life_saving_appliances" as "swimmer" and fails to detect certain objects under foggy and low-light conditions due to occlusion by light and fog. In contrast, the improved algorithm exhibits enhanced detection capabilities in foggy, low-light, and flare scenarios, effectively reducing both missed and false detections.



(**a**) Detection effect in foggy conditions

Figure 8. *Cont.*

(**b**) Detection effect in low-light conditions



(**c**) Detection effect in flare conditions

Figure 8. Examples of visualized object detection on the AWOD dataset, showing three typical severe weather conditions: (**a**) fog, (**b**) low light, and (**c**) flare. The first row represents ground truth, the second row represents RT-DETR results, and the third row shows our WRRT-DETR results.

Furthermore, due to the optimization of the multi-scale feature fusion network, the proposed algorithm significantly improves the detection of small-sized objects. The aspect ratios of the generated candidate bounding boxes are closer to the true bounding boxes. These results indicate that WRRT-DETR can detect objects more accurately, providing a robust solution for object detection tasks in complex scenarios.

6. Conclusions

In this study, AWOD has been introduced as the first large-scale dataset specifically designed for drone-view object detection under adverse weather conditions. The degradation of object detectors under such conditions has been observed and analyzed, leading to the development of WRRT-DETR to improve detection robustness. Additionally, the WRRT-DETR network has been constructed, incorporating the GLCE module to effectively leverage both local and global information. The FSAE has been introduced to facilitate the extraction of feature representations in complex environments. Furthermore, ACFM has been proposed to enhance small-object feature extraction while mitigating interference from complex background clutter. Experimental results have demonstrated that WRRT-DETR

exhibits significant advantages in small-object detection under challenging environmental conditions, effectively addressing the challenges posed by adverse weather. The findings of this study provide crucial support for drone-view object detection in complex weather conditions and establish a solid foundation for future research.

Author Contributions: Conceptualization, B.L.; methodology, B.L.; software, B.L.; validation, B.L. and C.S.; formal analysis, B.L.; investigation, B.L.; resources, B.L.; data curation, B.L.; writing—original draft preparation, B.L.; writing—review and editing, B.L. and C.S.; visualization, B.L.; supervision, Y.Z. and J.J.; project administration, Y.Z. and J.J.; funding acquisition, Y.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Shanghai Industrial Collaborative Innovation Project Foundation, grant number XTCX-KJ-2023-2-18, the Fundamental Research Funds for the Central Universities (Project 2232023D-27), and the Shanghai Pujiang Program (Project 23PJ1400300).

Data Availability Statement: We confirm that the data supporting the findings of this study are available from the following sources: [4,14–16]; the AWOD dataset is released at https://github.com/bei-liu/AWOD-datasets (accessed on 11 March 2025).

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

References

1. Zhang, Y.; Wu, C.; Guo, W.; Zhang, T.; Li, W. CFANet: Efficient detection of UAV image based on cross-layer feature aggregation. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5608911. [CrossRef]
2. Wang, K.; Fu, X.; Ge, C.; Cao, C.; Zha, Z.J. Towards generalized UAV object detection: A novel perspective from frequency domain disentanglement. *Int. J. Comput. Vis.* **2024**, *132*, 5410–5438. [CrossRef]
3. Wang, K.; Fu, X.; Huang, Y.; Cao, C.; Shi, G.; Zha, Z.J. Generalized uav object detection via frequency domain disentanglement. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 1064–1073.
4. Varga, L.A.; Kiefer, B.; Messmer, M.; Zell, A. Seadronessee: A maritime benchmark for detecting humans in open water. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2022; pp. 2260–2270.
5. Cafarelli, D.; Ciampi, L.; Vadicamo, L.; Gennaro, C.; Berton, A.; Paterni, M.; Benvenuti, C.; Passera, M.; Falchi, F. MOBDrone: A drone video dataset for man overboard rescue. In Proceedings of the International Conference on Image Analysis and Processing, Lecce, Italy, 23–27 May 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 633–644.
6. Liu, D.; Zhang, J.; Qi, Y.; Wu, Y.; Zhang, Y. Tiny object detection in remote sensing images based on object reconstruction and multiple receptive field adaptive feature enhancement. *IEEE Trans. Geosci. Remote. Sens.* **2024**, *62*, 5616213. [CrossRef]
7. Zhao, H.; Zhang, J.; Chen, Z.; Zhao, S.; Tao, D. Unimix: Towards domain adaptive and generalizable lidar semantic segmentation in adverse weather. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 14781–14791.
8. Testolina, P.; Barbato, F.; Michieli, U.; Giordani, M.; Zanuttigh, P.; Zorzi, M. Selma: Semantic large-scale multimodal acquisitions in variable weather, daytime and viewpoints. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 7012–7024. [CrossRef]
9. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef]
10. van Lier, M.; van Leeuwen, M.; van Manen, B.; Kampmeijer, L.; Boehrer, N. Evaluation of Spatio-Temporal Small Object Detection in Real-World Adverse Weather Conditions. In Proceedings of the Winter Conference on Applications of Computer Vision, Tucson, AZ, USA, 26 February–6 March 2025; pp. 844–855.
11. Gupta, H.; Kotlyar, O.; Andreasson, H.; Lilienthal, A.J. Robust object detection in challenging weather conditions. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2024; pp. 7523–7532.
12. Qi, Y.; Wang, D.; Xie, J.; Lu, K.; Wan, Y.; Fu, S. Birdseyeview: Aerial view dataset for object classification and detection. In Proceedings of the 2019 IEEE Globecom Workshops (GC Wkshps), Waikoloa, HI, USA, 9–13 December 2019; pp. 1–6.
13. Yu, X.; Gong, Y.; Jiang, N.; Ye, Q.; Han, Z. Scale match for tiny person detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Snowmass, CO, USA, 1–5 March 2020; pp. 1257–1265.
14. Du, D.; Zhu, P.; Wen, L.; Bian, X.; Lin, H.; Hu, Q.; Peng, T.; Zheng, J.; Wang, X.; Zhang, Y.; et al. VisDrone-DET2019: The vision meets drone object detection in image challenge results. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Republic of Korea, 27–28 October 2019.

15. 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]
16. Yu, F.; Chen, H.; Wang, X.; Xian, W.; Chen, Y.; Liu, F.; Madhavan, V.; Darrell, T. Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv* **2018**, arXiv:1805.04687.
17. Sakaridis, C.; Dai, D.; Van Gool, L. Semantic foggy scene understanding with synthetic data. *Int. J. Comput. Vis.* **2018**, *126*, 973–992. [CrossRef]
18. Qiu, Z.; Bai, H.; Chen, T. Special vehicle detection from UAV perspective via YOLO-GNS based deep learning network. *Drones* **2023**, *7*, 117. [CrossRef]
19. Bakirci, M. Real-time vehicle detection using YOLOv8-nano for intelligent transportation systems. *Trait. Signal* **2024**, *41*, 1727–1740. [CrossRef]
20. Hnewa, M.; Radha, H. Integrated multiscale domain adaptive YOLO. *IEEE Trans. Image Process.* **2023**, *32*, 1857–1867. [CrossRef]
21. Lin, H.; Li, Y.; Fu, X.; Ding, X.; Huang, Y.; Paisley, J. Rain o'er me: Synthesizing real rain to derain with data distillation. *IEEE Trans. Image Process.* **2020**, *29*, 7668–7680. [CrossRef]
22. Agyemang, S.A.; Shi, H.; Nie, X.; Asabere, N.Y. An integrated multi-scale context-aware network for efficient desnowing. *Eng. Appl. Artif. Intell.* **2025**, *151*, 110769. [CrossRef]
23. Zhu, Y.; Wang, T.; Fu, X.; Yang, X.; Guo, X.; Dai, J.; Qiao, Y.; Hu, X. Learning weather-general and weather-specific features for image restoration under multiple adverse weather conditions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 21747–21758.
24. Özdenizci, O.; Legenstein, R. Restoring vision in adverse weather conditions with patch-based denoising diffusion models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 10346–10357. [CrossRef] [PubMed]
25. Liang, D.; Li, L.; Wei, M.; Yang, S.; Zhang, L.; Yang, W.; Du, Y.; Zhou, H. Semantically contrastive learning for low-light image enhancement. In Proceedings of the AAAI Conference on Artificial Intelligence, Online, 22 February–1 March 2022; Volume 36, pp. 1555–1563.
26. Guo, X.; Hu, Q. Low-light image enhancement via breaking down the darkness. *Int. J. Comput. Vis.* **2023**, *131*, 48–66. [CrossRef]
27. Qin, Q.; Chang, K.; Huang, M.; Li, G. DENet: Detection-driven enhancement network for object detection under adverse weather conditions. In Proceedings of the Asian Conference on Computer Vision, Macao, China, 4–8 December 2022; pp. 2813–2829.
28. Sun, C.; Zhang, Y.; Ma, S. Dflm-yolo: A lightweight yolo model with multiscale feature fusion capabilities for open water aerial imagery. *Drones* **2024**, *8*, 400. [CrossRef]
29. Ma, S.; Zhang, Y.; Peng, L.; Sun, C.; Ding, L.; Zhu, Y. OWRT-DETR: A Novel Real-Time Transformer Network for Small Object Detection in Open Water Search and Rescue From UAV Aerial Imagery. *IEEE Trans. Geosci. Remote. Sens.* **2025**, *63*, 4205313. [CrossRef]
30. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
31. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8759–8768.
32. Tan, M.; Pang, R.; Le, Q.V. Efficientdet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10781–10790.
33. Tong, K.; Wu, Y. Small object detection using deep feature learning and feature fusion network. *Eng. Appl. Artif. Intell.* **2024**, *132*, 107931. [CrossRef]
34. Zhu, Z.; Zheng, R.; Qi, G.; Li, S.; Li, Y.; Gao, X. Small object detection method based on global multi-level perception and dynamic region aggregation. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 10011–10022. [CrossRef]
35. Zhao, H.; Zhang, H.; Zhao, Y. Yolov7-sea: Object detection of maritime uav images based on improved yolov7. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–7 January 2023; pp. 233–238.
36. Nautiyal, R.; Deshmukh, M. Tiny Object Detection for Marine Search and Rescue with YOLOv8n-Tiny. In Proceedings of the 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kamand, India, 24–28 June 2024; pp. 1–7.
37. Li, J.; Xu, R.; Ma, J.; Zou, Q.; Ma, J.; Yu, H. Domain adaptive object detection for autonomous driving under foggy weather. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–7 January 2023; pp. 612–622.
38. Marathe, A.; Ramanan, D.; Walambe, R.; Kotecha, K. Wedge: A multi-weather autonomous driving dataset built from generative vision-language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 3318–3327.
39. 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]

40. Zhao, J.; Chen, Y.; Zhou, Z.; Zhao, J.; Wang, S.; Chen, X. Multiship speed measurement method based on machine vision and drone images. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 2513112. [CrossRef]
41. Li, R.; Cheong, L.F.; Tan, R.T. Heavy rain image restoration: Integrating physics model and conditional adversarial learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 1633–1642.
42. Cai, Y.; Bian, H.; Lin, J.; Wang, H.; Timofte, R.; Zhang, Y. Retinexformer: One-stage retinex-based transformer for low-light image enhancement. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 12504–12513.
43. Ranftl, R.; Bochkovskiy, A.; Koltun, V. Vision transformers for dense prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 12179–12188.
44. Guo, T.; Li, S.; Zhou, Y.N.; Lu, W.D.; Yan, Y.; Wu, Y.A. Interspecies-chimera machine vision with polarimetry for real-time navigation and anti-glare pattern recognition. *Nat. Commun.* **2024**, *15*, 6731. [CrossRef] [PubMed]
45. Wang, X.; Liu, X.; Yang, H.; Wang, Z.; Wen, X.; He, X.; Qing, L.; Chen, H. Degradation Modeling for Restoration-enhanced Object Detection in Adverse Weather Scenes. *IEEE Trans. Intell. Veh.* **2024** . [CrossRef]
46. Yang, X.; Yan, J.; Liao, W.; Yang, X.; Tang, J.; He, T. Scrdet++: Detecting small, cluttered and rotated objects via instance-level feature denoising and rotation loss smoothing. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 2384–2399. [CrossRef]
47. Xu, C.; Ding, J.; Wang, J.; Yang, W.; Yu, H.; Yu, L.; Xia, G.S. Dynamic Coarse-To-Fine Learning for Oriented Tiny Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 7318–7328.
48. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. Detrs beat yolos on real-time object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 16965–16974.
49. Zhou, M.; Huang, J.; Yan, K.; Hong, D.; Jia, X.; Chanussot, J.; Li, C. A general spatial-frequency learning framework for multimodal image fusion. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024** . [CrossRef]
50. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
51. Wang, C.Y.; Yeh, I.H.; Mark Liao, H.Y. Yolov9: Learning what you want to learn using programmable gradient information. In Proceedings of the European Conference on Computer Vision, Milan, Italy, 29 September–4 October 2024; Springer: Berlin/Heidelberg, Germany, 2024; pp. 1–21.
52. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Ding, G. Yolov10: Real-time end-to-end object detection. *arXiv* **2024**, arXiv:2405.14458.
53. Khanam, R.; Hussain, M. Yolov11: An overview of the key architectural enhancements. *arXiv* **2024**, arXiv:2410.17725.
54. Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L.M.; Shum, H.Y. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv* **2022**, arXiv:2203.03605.
55. Liu, S.; Li, F.; Zhang, H.; Yang, X.; Qi, X.; Su, H.; Zhu, J.; Zhang, L. Dab-detr: Dynamic anchor boxes are better queries for detr. *arXiv* **2022**, arXiv:2201.12329.
56. Li, F.; Zhang, H.; Liu, S.; Guo, J.; Ni, L.M.; Zhang, L. Dn-detr: Accelerate detr training by introducing query denoising. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 13619–13627.
57. Feng, C.; Zhong, Y.; Gao, Y.; Scott, M.R.; Huang, W. Tood: Task-aligned one-stage object detection. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 3490–3499.
58. Xu, J.; Fan, X.; Jian, H.; Xu, C.; Bei, W.; Ge, Q.; Zhao, T. Yoloow: A spatial scale adaptive real-time object detection neural network for open water search and rescue from uav aerial imagery. *IEEE Trans. Geosci. Remote. Sens.* **2024**, *62*, 5623115. [CrossRef]
59. Shen, H.; Lin, D.; Song, T. Object detection deployed on UAVs for oblique images by fusing IMU information. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 6505305. [CrossRef]