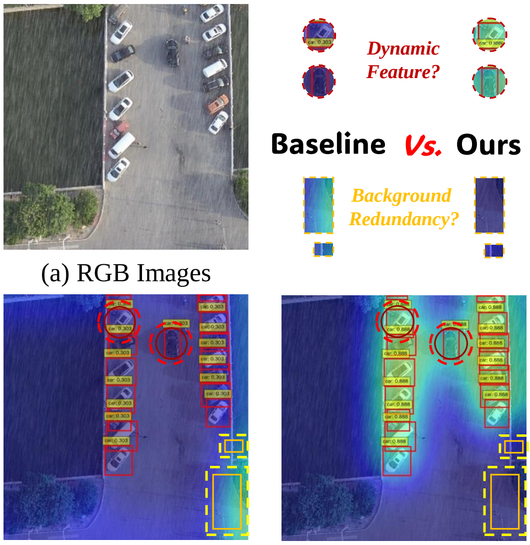
Learning Geometric-Aware and Weather-Adaptive Semantics in Remote Sensing: Affine Lie Group Enhanced Detector for UAV Road Scenes

Jialang Liu, Jialei Zhan, Yanming Guo, Taiyong Li, Yimeng Zhao, Jiehua Zhang, Lixing Tang, Yishan Li, Yingmei Wei, Weiwei Cai, *Member, IEEE*

***Abstract*—Reliable road condition detection using drone imagery is critically important, particularly under harsh weather conditions such as rain, fog, and snow, which cause reduced visibility and blurred objects. Traditional detection methods are limited in effectively handling these severe scenarios due to their static feature extraction approaches. To address these challenges, we propose an innovative affine lie group convolution and weather-adaptive feature enhancement network (ALGC-WFEMNet). The core innovation of this method lies in the affine lie group convolution (ALGC), which leverages the mathematical framework of affine lie groups to introduce a dynamic convolution mechanism. This mechanism adaptively modifies convolution kernels based on affine transformations, significantly enhancing the model's robustness against weather-induced variations in scale, rotation, and visibility. Furthermore, the ALGC framework integrates a learning-based weather condition coefficient, dynamically adjusting kernel responses to specific environmental conditions such as rain, fog, and snow. This theoretical advancement not only emphasizes the mathematical novelty of applying affine lie groups in convolutional neural networks but also substantially improves feature extraction and adaptability for object detection tasks. Experimental validation on UAV-based road inspection datasets demonstrates that our ALGC-WFEMNet achieves a mean average precision (mAP) of 60.48%. Furthermore, we deploy the model within a UAV-IoT system to verify its practical effectiveness, achieving an inference time of 23.31 seconds on a Raspberry Pi.**

***Index Terms*—remote sensing, UAV, object detection, Lie Group, harsh weather**



(a) RGB Images

(b) Baseline (c) ALGC-WFEMNet

Fig. 1. Visual comparison of feature attribution under adverse weather. (a) Input UAV RGB images captured in rain, fog and snow. (b) Baseline model (Cascade R-CNN + MobileSAM) heatmaps, showing dispersed and background-prone attention. (c) ALGC-WFEMNet heatmaps, highlighting improved focus on semantic targets and suppression of background noise.

Remote Sensing Object Detection (RSOD) is essential for applications such as environmental monitoring, military surveillance, and urban management. In drone-based aerial imaging, however, detection accuracy often degrades under adverse weather. Rain, haze, turbulence, and low illumination obscure object geometry and weaken key cues such as contours, textures, and color contrast, causing object shapes to blend into the background. These degradations introduce not only appearance noise but also geometric distortion, making it difficult for conventional detectors to maintain stable performance. In scenarios such as highway vehicle monitoring, for example, wet surfaces or low visibility can alter object boundaries and suppress structural cues, which undermines both recognition and localization. These effects highlight a fundamental challenge: weather-induced degradations and geometric variations interact closely, yet most existing methods treat them in isolation. A unified modeling framework that jointly accounts for geometric transformations and weather-related feature degradation is therefore critical for achieving robust object detection in real-world remote sensing environments.

Detecting small objects, such as vehicles on highways, in remote sensing images presents an additional challenge: the objects occupy only a small number of pixels, and their feature responses are weak, particularly in complex backgrounds. This

This work was supported by the National Natural Science Foundation of China (General Program, Grant No. 72571281), the Excellent Young Scientist Fund of Hunan Province (Grant No. 2025JJ40066), and the Training Program for Excellent Young Innovators of Changsha (Grant No. kq2209001). Corresponding authors: Yanming Guo and Weiwei Cai. Jialang Liu, Jialei Zhan and Yanming Guo contributed equally to this work.

Jialang Liu, Jialei Zhan, Yimeng Zhao, Yishan Li, Yanming Guo, and Yingmei Wei are with the Laboratory for Big Data and Decision, National University of Defense Technology(NUDT), Changsha 410004, China (email: liu_1999@nudt.edu.cn; jieleiz@163.com; guoyanming@nudt.edu.cn; zhaoym20@nudt.edu.cn; liyishan@nudt.edu.cn; weiyingmei@nudt.edu.cn). Taiyong Li is with the College of Electronic Information and Physics, Central South University of Forestry and Technology, Changsha 410004, China (email: taiyyyyyli@163.com, 20233702@csuft.edu.cn). Jiehua Zhang is with the Center for Machine Vision and Signal Analysis, University of Oulu, Finland (email: jiehua.zhang@oulu.fi). Weiwei Cai (Member, IEEE) is with the School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi 214122, China(email: vivitsai@ieee.org).

issue is especially pronounced in aerial images of multi-lane highways. Feature Pyramid Networks (FPNs), which are widely used in object detection, address the challenge of scale variation by constructing multi-level feature pyramids using deep neural networks. FPNs combine high-resolution detail information with low-resolution global context through top-down feature fusion and lateral connections, making them effective for detecting both large objects near the camera and small, distant objects, such as vehicles on highways.

In adverse weather conditions, as illustrated in Figure 1, traditional feature extractors often struggle to distinguish foreground targets from cluttered or low-contrast backgrounds. To address the limitations of conventional multi-scale fusion under such conditions, we propose an innovative affine lie group convolution and weather-adaptive feature enhancement network (ALGC-WFEMNet), a novel architecture specifically designed for UAV-based object detection in adverse weather. This framework enhances the representational power of the Feature Pyramid Network (FPN) through the integration of Affine Lie Group Convolution (ALGC) in its feature fusion layers, enabling better modeling of geometric variations and improving sensitivity to small and ambiguous targets. Furthermore, we introduce the Weather-Adaptive Feature Enhancement Module (WFEM), which dynamically amplifies inter-channel feature distinctions and restores spatial-semantic relationships disrupted by environmental interference. By providing adaptive priors to the ALGC module, WFEM ensures more reliable feature modulation in scenarios involving haze, rain, or non-uniform lighting, ultimately boosting detection robustness and accuracy.

Our contributions are summarized as follows:

- We construct a custom dataset for drone-based highway object detection under adverse weather conditions. To address the challenges of object detection in harsh weather environments, we develop a large-scale, high-quality dataset specifically tailored for drone-based highway scenarios. This dataset includes diverse weather conditions such as rain, snow, and fog, and features a wide range of vehicle types and scales. The dataset is meticulously annotated to ensure accuracy and consistency, providing a robust foundation for training and evaluating object detection models under adverse weather conditions.
- We propose a novel FPN structure specifically designed for drone-based road detection in harsh weather. Our ALGC-WFEMNet framework introduces significant improvements to the traditional FPN by incorporating modules specifically designed to address the challenges posed by adverse weather. This novel structure enhances the network's ability to capture multi-scale information and adapt to the unique characteristics of drone-based road imagery, making it particularly effective for detecting objects in complex and dynamic environments.
- We introduce a novel convolution operator, Affine Lie Group Convolution (ALGC), which explicitly integrates affine transformations into convolutional operations through Lie group theory. The ALGC module dynamically predicts affine transformation parameters directly from input features, enabling the model to robustly adapt to variations in object scale, rotation, and orientation caused by adverse weather conditions such as fog, rain, and snow. By leveraging geometric transformations explicitly encoded through Lie algebra exponential mappings, ALGC significantly improves the model's ability to detect objects accurately and reliably, even when visibility is severely impaired or objects appear distorted due to environmental challenges.
- We propose a lightweight module, WFEM, which requires minimal additional computation and memory, to highlight pixel-level feature relationships, reduce noise, and improve robustness against adverse weather. WFEM leverages a dual-branch structure to modulate feature maps at the pixel level, effectively suppressing noise and enhancing discriminative features. This lightweight module is computationally efficient and memory-friendly, making it suitable for real-time applications. By improving the clarity and robustness of feature representations, WFEM ensures reliable object detection performance in a wide range of adverse weather scenarios.

I. Related Work

Recently, there have been many developments in lightweight object detection work. This section briefly reviews the work of existing lightweight remote sensing object detection and binarized neural networks used for general object detection.

A. Remote sensing Object Detection

Early research on RSOD relied heavily on single-scale feature maps produced by the final backbone layer. Although simple, these representations were unable to cope with the wide range of object scales found in remote sensing images [1]–[3]. This limitation gradually pushed the community toward multi-scale feature learning, and related studies have mainly followed three interconnected directions: multi-level feature fusion, pyramid feature hierarchies, and feature pyramid networks.

The first line of work seeks to fuse features from different depths to obtain a richer representation. Shallow layers preserve spatial details such as edges and textures, whereas deeper layers contain more abstract semantic cues. By combining them, the fused representation becomes more suitable for detecting objects of diverse sizes. Representative efforts include hierarchical fusion of multiple convolutional levels [4], normalization-based fusion strategies to reduce scale mismatch across layers [5], and the construction of multi-receptive-field features using atrous separable convolutions within a single layer [6], [7]. These works established the foundation for integrating complementary information across layers. The second direction explores the idea of independently detecting objects at different feature levels. This perspective was popularized by SSD [8], which arranges prediction heads at multiple depths so that small objects are handled by high-resolution layers while larger ones are addressed by deeper layers. Variants designed for remote sensing have further enhanced this paradigm. Examples include the addition of dedicated branches for small vehicles [9], scale-invariant regression layers that jointly supervise multiple depths [10], and hierarchical filtering modules using multi-size kernels to extract multi-receptive-field features [11]. These methods reinforced the idea that scale diversity can be explicitly handled by distributing predictions across layers. The third direction evolves from the observation that multi-layer predictions alone cannot fully exploit the complementary nature of shallow and deep features. Feature Pyramid Networks (FPN) [12] introduced a top–down pathway to strengthen the semantic content of high-resolution features, and this architecture has since become a central component in multi-scale modeling. Subsequent studies refined the framework by improving its suitability for remote sensing scenes: asymmetric convolutional structures were introduced to capture elongated objects such as bridges and runways [13], high-frequency details were injected to preserve structural cues [14], and bidirectional fusion schemes were developed to compensate for semantic loss during long-distance propagation within deep backbones. Layer-wise attention mechanisms further enhanced fusion quality by learning the relative importance of different levels [15].

Although these multi-scale strategies have substantially improved detection accuracy, their adaptability is still limited when faced with adverse weather, occlusion, noise, or extremely unbalanced object scales. To address these issues, we introduce an Affine Lie Group Convolution (ALGC) module. ALGC incorporates affine geometric priors together with environment-aware modulation, enabling the convolution kernels to adapt to contextual variations while retaining spatial structures and semantic cues. As demonstrated in our experiments, this design significantly enhances model robustness and detection reliability across challenging remote sensing environments [16].

B. Object Detection in Inclement Weather Conditions

Object detection in adverse weather conditions presents a critical challenge for autonomous driving systems, as environmental factors such as fog, rain, snow, and low-light conditions can significantly degrade detection accuracy and robustness. These challenges have spurred substantial research focused on mitigating visibility degradation, sensor noise, and domain shifts. Existing methods can be broadly categorized into four main strategies: model architecture improvements, uncertainty estimation, image enhancement, and domain adaptation.

Model architecture improvements focus on enhancing detection frameworks to better handle adverse weather. For instance, anchor-free designs and decoupled detection heads have been integrated into YOLOv4 to improve multi-scale detection accuracy and speed [17], [18]. Similarly, dual-subnet networks (e.g., DSNet) combine visibility enhancement with object detection to improve performance in foggy conditions [19]. YOLOv5-based models have also been optimized for diverse weather scenarios, incorporating advanced modules like Transformers and CBAM to enhance feature extraction [20], [21]. However, these methods often rely on specific architectural modifications tailored to certain weather conditions, limiting their generalizability across diverse and extreme weather scenarios. Uncertainty estimation methods, such as Bayesian approaches, assess prediction reliability under adverse conditions [22]. These models introduce metrics like anomaly detection ratios to evaluate detection confidence, particularly in scenarios such as nighttime driving or snow. While effective in quantifying uncertainty, these methods do not directly address the degradation of input data quality, which is essential for robust detection performance. Image enhancement techniques aim to improve the quality of input data for better detection. Methods like color-level shift compensation [21] and image-adaptive YOLO (IA-YOLO) [23] adaptively enhance images to improve clarity and detection performance. Additionally, polarization imaging has been explored to leverage multi-dimensional information for improved detection accuracy under adverse weather [24]. However, these methods often focus on specific weather conditions (e.g., fog or low light) and require additional computational resources or specialized sensors, limiting their scalability and applicability in real-world scenarios. Domain adaptation strategies address the domain gap between clear and adverse weather images. For example, unsupervised domain adaptation frameworks decompose the domain gap into style and weather factors, using attention modules and contrastive learning to improve robustness [25]. While these methods effectively handle domain shifts, they often require extensive training on diverse datasets and do not fully exploit the complementary information across different weather conditions.

Despite these advancements, existing methods often target specific weather conditions or challenges, such as fog or domain shifts, without providing a unified solution for diverse and extreme scenarios. Furthermore, their computational complexity and reliance on specialized sensors (e.g., polarization imaging) limit scalability and real-time applicability. To overcome these limitations, we propose the Weather-Adaptive Feature Enhancement Module (WFEM), which improves detection performance under various adverse weather conditions by leveraging dynamic feature fusion, weather-aware adjustment, and multi-scale adaptation strategies [26], [27], [28].

II. ALGC-WFEMNET

In this section, we provide a detailed description of our ALGC-WFEMNet (as shown in Figure 2). For clarity, the abbreviations appearing in Fig. 2 follow the module naming in our framework, where BTNK denotes the bottleneck block, WFEM refers to the Weather-Adaptive Feature Enhancement Module, and ALGC represents the proposed Affine Lie Group Convolution. All of our improvements are implemented on the FPN. We introduce our ALGC-WFEMNet, which is designed to capture the rich multi-scale information within the FPN. Subsequently, we describe our Weather-Adaptive Feature Enhancement Module (WFEM), a method that removes redundant information and enhances the model's resistance to interference.

A. Baseline: Cascade R-CNN and MobileSAM

Cascade R-CNN and MobileSAM have emerged as prominent baseline methods in contemporary object detection research due to their effectiveness and efficiency [29]. In particular, Cascade R-CNN serves as a strong high-accuracy benchmark for evaluating improvements in multi-scale geometric modeling, which is essential in adverse-weather UAV imagery. Cascade R-CNN, built upon the DETR framework, introduces innovative improvements such as denoising training and contrastive queries, effectively enhancing detection accuracy and convergence speed, especially in complex scenes with dense object distributions [30]. Its transformer-based architecture naturally captures global contextual relationships, significantly boosting detection robustness across varied scenarios.

On the other hand, MobileSAM, a streamlined and lightweight variant derived from the Segment Anything Model (SAM), provides superior segmentation capabilities tailored explicitly for mobile and resource-constrained environments [31]. Because MobileSAM excels at generating clean and reliable region boundaries even with limited computation, it offers an ideal segmentation prior for analyzing how weather-induced degradations influence boundary integrity and object spatial coherence. Despite its compactness, MobileSAM maintains commendable segmentation quality, facilitating accurate boundary delineation and region proposal generation with considerably reduced computational overhead [32]. Its efficiency also aligns with real UAV deployment settings, where on-board resources and real-time constraints limit the feasibility of heavier segmentation pipelines.

By combining Cascade R-CNN's refined object localization ability with MobileSAM's efficient segmentation approach, these methods serve as strong baselines, demonstrating high potential for real-world deployment in object detection tasks, particularly when computational resources and inference latency pose critical constraints. Together, they provide complementary viewpoints, one emphasizing high-precision geometric reasoning and the other emphasizing weather-sensitive boundary cues, allowing a more comprehensive validation of our proposed approach.

B. Lie Group $\mathrm{SO}(2)$ *and Its Potential in Convolutional Architectures.*

The special orthogonal group $\mathrm{SO}(2)$ represents all two-dimensional rotations that preserve the origin and the Euclidean norm. Formally, it is defined as

$$\mathrm{SO}(2) = \left\{ R(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \;\middle|\; \theta \in [0, 2\pi) \right\}, \quad (1)$$

where each $R(\theta)$ is a 2×2 real orthogonal matrix with determinant equal to one. The group $\mathrm{SO}(2)$ is continuous, compact, and forms a one-dimensional Lie group, with group operation defined by matrix multiplication.

Integrating the structure of $\mathrm{SO}(2)$ into convolutional neural networks offers a promising way to achieve rotation-equivariant representations. Conventional convolutions are naturally equivariant to translations but not to rotations, which can lead to performance degradation when objects appear at arbitrary orientations. By embedding $\mathrm{SO}(2)$ symmetry into the design of convolutional layers, such as through group convolutions or Lie group parameterizations, models can become intrinsically sensitive to rotated patterns without requiring extensive data augmentation. This design enhances robustness and data efficiency, especially in vision tasks where rotational variability is common, including remote sensing, medical imaging, and autonomous navigation.

C. Affine Lie Group Convolution (ALGC)

Adverse weather conditions such as rain, fog, and snow introduce complex geometric and photometric distortions to aerial imagery, which severely compromise the robustness of standard convolutional neural networks (CNNs). In particular, scale variance, rotation, and partial occlusions break the spatial stationarity assumption underlying traditional convolution operations, leading to degraded feature extraction and weakened generalization.

To overcome this, we introduce a novel **Affine Lie Group Convolution (ALGC)** module, which explicitly incorporates the mathematical structure of affine transformations into convolutional feature extraction. By embedding affine Lie group actions into the convolution process, ALGC enables spatially adaptive and weather-aware filtering, improving robustness to geometry-induced feature variation. The core idea is to dynamically predict and apply affine transformations on feature maps before performing spatial convolution, thus endowing the network with geometric flexibility in feature encoding.

Affine parameter prediction. Given an input feature map $x_{\text{in}} \in \mathbb{R}^{B\times C\times H\times W}$, where B, C, H, and W denote the batch size, number of channels, height, and width respectively, we first extract a compact global descriptor through spatial average pooling. This descriptor is then passed through a 1×1 convolution layer to regress six affine parameters for each sample in the batch:

$$\theta = \mathrm{Conv}_{1\times 1}(\mathrm{AvgPool}(x_{\text{in}})) \in \mathbb{R}^{B\times 6} \quad (2)$$

This lightweight parameterization predicts an affine transformation matrix of the form:

$$\hat{\theta} = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} \in \mathbb{R}^{2\times 3} \quad (3)$$

which governs scaling, rotation, shear, and translation operations.

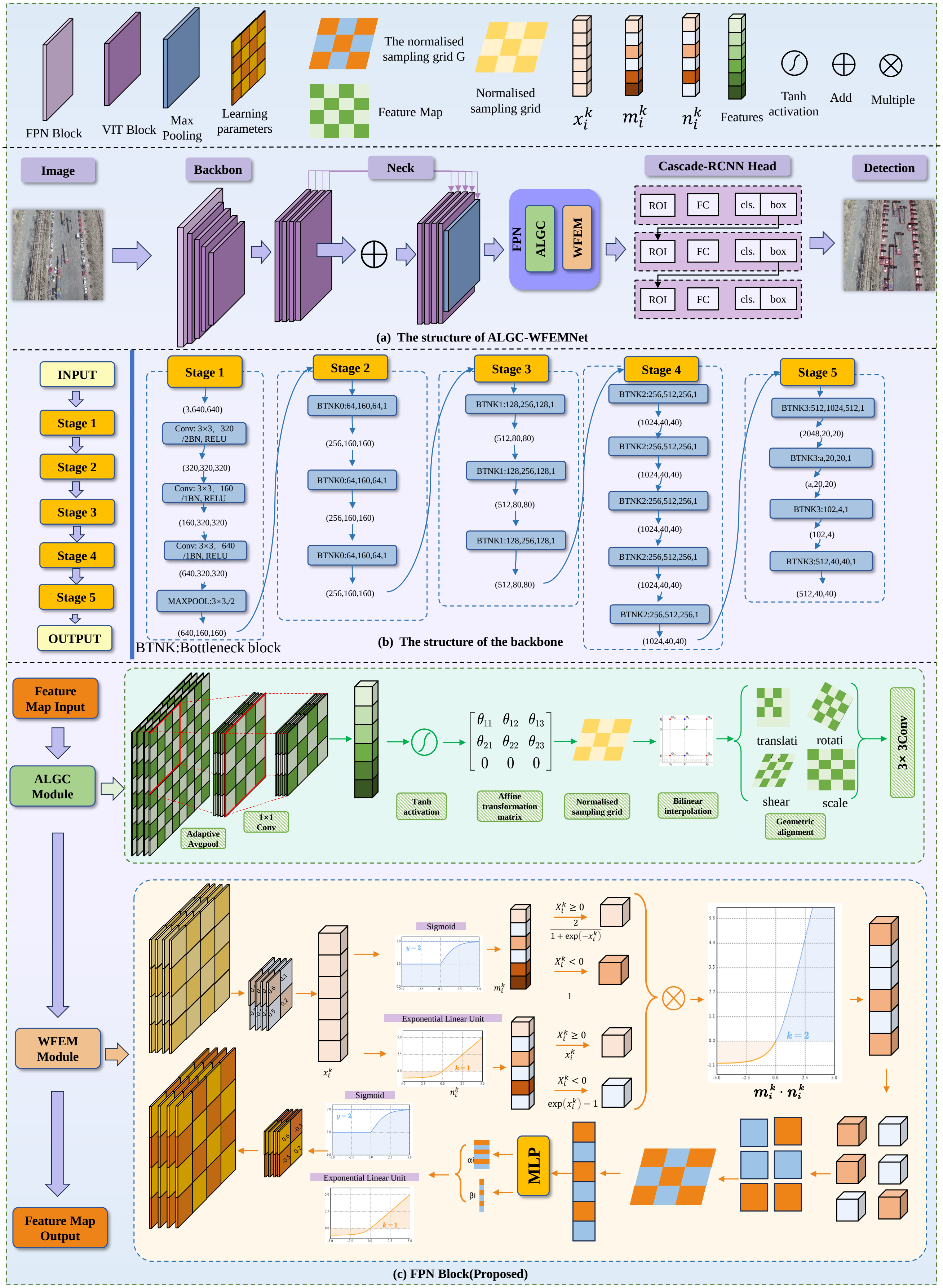
To maintain numerical stability during training and inference, we constrain the predicted values via a hyperbolic tangent activation, scaled by a small constant factor λ:

$$\theta = \lambda \cdot \tanh(\theta), \quad \lambda = 0.1 \quad (4)$$

This regularization ensures the predicted affine transformations remain within a moderate range, avoiding excessive warping or instability during gradient descent optimization. The value $\lambda = 0.1$ was empirically selected to balance expressiveness and smoothness of the spatial transformation.

After normalization, the parameter vector θ is reshaped into a batched affine matrix format:

$$\theta \;\longrightarrow\; \hat{\theta} \in \mathbb{R}^{B\times 2\times 3} \quad (5)$$



(a) The structure of ALGC-WFEMNet

(b) The structure of the backbone

(c) FPN Block(Proposed)

Fig. 2. The architecture of ALGC-WFEMNet and its key components. (a) The overall structure of ALGC-WFEMNet, where the backbone extracts multi-scale features (S2, S3, S4, and S5 represent feature extraction stages, where larger values correspond to deeper feature layers with stronger semantic information), and the proposed FPN blocks (green) enhance feature fusion for robust object detection under adverse weather conditions. (b) The proposed FPN block integrates the Weather-Adaptive Feature Enhancement Module (WFEM) and Affine Lie Group Convolution (ALGC) to improve feature representation and adaptability. (c) The WFEM module employs a dual-branch structure to modulate features, suppress noise, and enhance robustness against challenging weather conditions.

Feature warping via affine group action. With the affine transformation matrices in hand, we perform a differentiable spatial warping on the input features. This warping simulates the geometric action of the affine Lie group $\mathrm{Aff}(2)$ on the input grid domain. Specifically, we generate a normalized sampling grid G based on the affine parameters:

$$G = \mathrm{AffineGrid}(\hat{\theta}, \mathrm{size}(x_{\mathrm{in}}), \mathrm{align\_corners{=}True}) \tag{6}$$

We then use bilinear interpolation to warp the original feature map along the learned transformation grid:

$$\begin{aligned} x_{\mathrm{warp}} = \mathrm{GridSample}(x_{\mathrm{in}}, G, \ & \mathrm{mode} = \mathrm{bilinear}, \\ & \mathrm{padding\_mode} = \mathrm{border}, \\ & \mathrm{align\_corners} = \mathrm{True}) \end{aligned} \tag{7}$$

this operation aligns distorted objects to canonical geometry, facilitating consistent feature extraction, compensating for adverse distortions and normalizing the spatial variation caused by environmental factors.

Convolution over warped features. Once the input has been transformed via affine warping, we apply a standard $k \times k$ convolution over the geometrically normalized representation:

$$x_{\mathrm{out}} = \mathrm{Conv}_{k\times k}(x_{\mathrm{warp}}; w, b), \quad k = 3 \tag{8}$$

This final step produces a spatially adaptive feature representation that integrates both global semantics and local geometry, enhancing detection sensitivity to small and distorted objects in complex weather.

Lie group perspective. The affine transformation matrices used in ALGC form a subgroup of the general linear group, and can be viewed as elements of the affine Lie group $\mathrm{Aff}(2)$, which is a six-dimensional, non-compact, non-abelian Lie group. In principle, any transformation matrix $T \in \mathrm{Aff}(2)$ can be written as an exponential map from a corresponding Lie algebra element $\mathfrak{g} \in \mathbb{R}^{2\times 3}$:

$$T = \exp\left(\begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \\ 0 & 0 & 0 \end{bmatrix}\right) \tag{9}$$

While our implementation directly regresses the affine parameters $\hat{\theta}$ for practical reasons, the underlying formulation adheres to the geometric structure of Lie groups and enables potential extensions to more general transformation spaces (e.g., projective groups or conformal groups). This connection bridges group-theoretic insights with neural network design, paving the way for principled equivariant learning under affine actions.

D. Weather-Adaptive Feature Enhancement Module (WFEM)

Under adverse weather conditions, such as rain, fog, and snow, the model's ability to represent local features is significantly weakened, and higher-order information is lost [33]. This degradation is caused by the challenging environmental factors, which reduce the clarity and discriminative power of the extracted features, ultimately impacting detection performance [34]. With this in mind, we propose Weather-Adaptive Feature Enhancement Module (WFEM), which modulates the input features to serve the 1×1 convolution and adopts the modulated features through the fusion form.

WFEM is designed as two branches. The first branch is used to compute modulation coefficients from the input feature map that acts on the positive half. The second branch directly modulates the input feature map to suppress the negative half while keeping the positive half unchanged. Then, the two branches are merged using the Hadamard product to obtain the modulation result. For the first branch, the Sigmoid function with smoother output is chosen. It is vertically shifted, and its multiplication is expanded to ensure an appropriate range of modulation coefficients, prevent the original relationship between feature values from being broken, and enhance the distinction between different pixels.

$$m_i^k = \begin{cases} \frac{2}{1+\exp\left(-x_i^k\right)} \text{ if } x_i^k \geq 0 \\ 1 \text{ if } x_i^k < 0 \end{cases} \tag{10}$$

where x_i^k represents the i-th pixel point of the feature map, and k is the channel index of the feature map, and m_i^k represents the output of the first branch.

For the second branch, although the convolution results in a weaker representation of the output feature map, even so, negative feature values still contain semantic information. For this reason, we use the elu function with its non-linear fitting properties to suppress negative feature values in the form of an exponent on the negative half-axis and keep the feature values constantly on the positive half-axis.

$$n_i^k = \begin{cases} x_i^k \text{ if } x_i^k \geq 0 \\ \exp\left(x_i^k\right) - 1 \text{ if } x_i^k < 0 \end{cases} \tag{11}$$

where x_i^k represents the i-th pixel point of the feature map, and k is the channel index of the feature map, and n_i^k represents the output of the second branch. Next, we use the Hadamard product of the two branches above to fuse the output features of the two branches, and the fusion process can be expressed as follows:

$$m' * n' = m_i^k n_i^k = \begin{cases} \frac{2x_i^k}{1+\exp\left(-x_i^k\right)} \text{ if } x_i^k \geq 0 \\ \exp\left(x_i^k\right) - 1 \text{ if } x_i^k < 0 \end{cases} \tag{12}$$

m' and n' represent the output feature maps of the first and second branches, respectively.

WFEM demonstrates strong adaptability to various adverse weather conditions through its dual-branch feature enhancement strategy. In rainy conditions, raindrops and streaks cause blurring and occlusion, while WFEM's channel attention mechanism amplifies high-frequency edge details, suppressing rain noise and enhancing contours to mitigate blur effects. In snowy conditions, reduced contrast causes objects to blend into the background, leading to low-contrast distortion. WFEM employs a Sigmoid modulation function to enhance feature contrast and reduce confusion. In foggy conditions, light scattering and contrast loss blur object boundaries. WFEM applies global channel weighting to strengthen object responses, preserving structural integrity. Overall, WFEM dynamically adjusts feature mappings to counteract feature degradation under adverse weather conditions.

To compactly express the entire WFEM operation, we formulate it as an element-wise modulation over the input feature map $\mathbf{X} \in \mathbb{R}^{B\times C\times H\times W}$. The dual-branch fusion process is summarized as:

$$\mathbf{Y} = \text{WFEM}(\mathbf{X}) = \sigma_{\text{pos}}(\mathbf{X}) \odot \phi_{\text{neg}}(\mathbf{X}) \tag{13}$$

where $\sigma_{\text{pos}}(\cdot)$ is a channel-wise, positively modulated activation branch defined by:

$$\sigma_{\text{pos}}(x) = \begin{cases} \frac{2}{1+\exp(-x)} & x \geq 0 \\ 1 & x < 0 \end{cases}, \tag{14}$$

and $\phi_{\text{neg}}(\cdot)$ is a negative-preserving enhancement branch defined by:

$$\phi_{\text{neg}}(x) = \begin{cases} x & x \geq 0 \\ \exp(x) - 1 & x < 0 \end{cases}. \tag{15}$$

The final enhanced feature $\mathbf{Y}$ preserves positive activations while smoothly suppressing and refining negative regions in a data-dependent manner. This unified formulation emphasizes the element-wise, nonlinear modulation behavior of WFEM at the tensor level, providing a compact yet expressive characterization of its functional role in robust feature refinement under adverse weather conditions.

To further improve the environmental adaptability of the enhancement process, we extend WFEM by incorporating a lightweight graph-based context modeling mechanism and a reinforcement learning-guided modulation policy. This enables the network to dynamically adjust the feature modulation strategy according to the spatially variant degradation characteristics induced by different weather conditions. Specifically, the input feature map $\mathbf{X} \in \mathbb{R}^{B\times C\times H\times W}$ is first divided into N non-overlapping patches $\{\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_N\}$, each corresponding to a node in the degradation graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. For each patch $\mathbf{P}_i$, we extract a local degradation descriptor $\mathbf{d}_i \in \mathbb{R}^D$ capturing low-level cues such as average intensity, local contrast variance, gradient entropy, and haze level estimates.

The node descriptors are used to initialize the graph node embeddings $\mathbf{v}_i^{(0)} = \mathbf{d}_i$, and a multi-head Graph Attention Network (GAT) is applied to propagate contextual information across connected regions. The graph update process is defined as:

$$\mathbf{v}_i^{(l+1)} = \sigma\left(\sum_{j\in\mathcal{N}(i)} \alpha_{ij}^{(l)} \mathbf{W}^{(l)} \mathbf{v}_j^{(l)}\right), \tag{16}$$

where $\alpha_{ij}^{(l)}$ is the attention coefficient between nodes i and j in layer l, computed as:

$$\alpha_{ij}^{(l)} = \frac{\exp\left(\text{LeakyReLU}(\mathbf{a}^\top[\mathbf{W}^{(l)}\mathbf{v}_i^{(l)} \parallel \mathbf{W}^{(l)}\mathbf{v}_j^{(l)}])\right)}{\sum_{k\in\mathcal{N}(i)} \exp\left(\text{LeakyReLU}(\mathbf{a}^\top[\mathbf{W}^{(l)}\mathbf{v}_i^{(l)} \parallel \mathbf{W}^{(l)}\mathbf{v}_k^{(l)}])\right)}, \tag{17}$$

and $\mathbf{W}^{(l)}$, $\mathbf{a}$ are learnable parameters. The final output $\mathbf{z}_i = \mathbf{v}_i^{(L)}$ encodes the weather-aware representation of patch i.

These contextual embeddings are passed into a policy network π_θ to predict the adaptive modulation parameters α_i, β_i for the corresponding patch. Formally,

$$(\alpha_i, \beta_i) = \pi_\theta(\mathbf{z}_i), \tag{18}$$

where π_θ is a shallow MLP trained via reinforcement learning. The environment state is defined by the degradation descriptor $\mathbf{z}_i$, and the action space consists of continuous-valued (α, β) pairs that control the shape of the modulation functions. The reward signal $\mathcal{R}_i$ is defined to encourage spatial consistency, confidence boost, and reduced entropy in the modulated features:

$$\mathcal{R}_i = \lambda_1 \cdot \Delta\text{IoU}_{\text{local}} + \lambda_2 \cdot \Delta\text{Conf}_{\text{avg}} - \lambda_3 \cdot \mathcal{H}(\mathbf{Y}_i), \tag{19}$$

where $\Delta\text{IoU}_{\text{local}}$ measures detection improvement in region i, $\Delta\text{Conf}_{\text{avg}}$ is the average confidence change before and after modulation, and $\mathcal{H}(\mathbf{Y}_i)$ denotes the entropy of the enhanced region.

Using the predicted parameters, we generalize the original dual-branch modulation to dynamic forms. The positive activation function becomes:

$$\sigma_{\text{pos}}^{\alpha,\beta}(x) = \begin{cases} \frac{\alpha}{1+\exp(-\beta x)} & x \geq 0 \\ 1 & x < 0 \end{cases}, \tag{20}$$

and the negative enhancement function becomes:

$$\phi_{\text{neg}}^{\beta}(x) = \begin{cases} x & x \geq 0 \\ \exp(\beta x) - 1 & x < 0 \end{cases}. \tag{21}$$

The final enhanced feature map is computed element-wise as:

$$\mathbf{Y}_i = \sigma_{\text{pos}}^{\alpha_i,\beta_i}(\mathbf{X}_i) \odot \phi_{\text{neg}}^{\beta_i}(\mathbf{X}_i), \tag{22}$$

where $\mathbf{X}_i$ is the feature patch corresponding to node i, and $\mathbf{Y}_i$ is its enhanced output. In this way, the original WFEM formulation is extended into a unified dynamic modulation framework guided by global weather context and learned enhancement policies. This design significantly boosts the robustness of the model against spatially varying visual degradations and enables policy-driven modulation decisions that are optimized jointly with detection objectives.

III. Experiments

A. Datasets

To address the challenges of object detection in adverse weather conditions, we constructed a comprehensive dataset, the Inclement-weather UAV-based Multi-class highway dataset (IMC), consisting of 36,390 aerial images captured using a DJI Mavic 3 drone, which is equipped with a high-resolution camera capable of capturing detailed aerial imagery. The data collection process was conducted across diverse urban and suburban environments, including intersections, highways, and parking lots, to ensure a wide range of traffic scenarios. The drone was operated at varying altitudes between 50 and 150 meters, balancing spatial coverage and object-level detail.

To simulate adverse weather conditions, the dataset includes images captured during natural weather events such as rain and snow, as well as images augmented with synthetic weather effects to enhance robustness. The dataset focuses on five distinct vehicle categories: car, truck, bus, van, and freight car, ensuring a diverse representation of real-world traffic patterns. The annotation process was carried out using the open-source tool labelImg. Each image was manually labeled by a team of trained annotators to ensure high-quality annotations. The process involved drawing precise bounding boxes around each object of interest and assigning one of the five predefined class labels to each bounding box. To maintain consistency and accuracy, a multi-stage quality control process was implemented, including cross-validation by multiple annotators and periodic reviews by senior annotators. The final dataset contains high-quality annotations with minimal noise, making it suitable for training and evaluating object detection models under adverse weather conditions.

Moreover, Fig. 3 shows some typical scenes in the IMC dataset, which covers a variety of adverse weather scenarios such as snow, fog, rain, and abnormal lighting conditions.

The dataset was divided into training, validation, and testing subsets to facilitate model development and evaluation. Specifically, 70% of the images (25,473) were allocated to the training set, 15% (5,459) to the validation set, and the remaining 15% (5,458) to the testing set. The division was performed randomly while ensuring that each subset maintained a similar distribution of weather conditions and object categories to prevent data imbalance. This dataset provides a valuable resource for advancing research in adverse weather object detection, with potential applications in autonomous driving systems, traffic monitoring, and UAV-based surveillance in challenging environments.

It is worth noting that in order to evaluate the performance of the model under different weather conditions, we constructed a dataset containing both real and synthetic weather images. This dataset contains 20,000 real weather images (captured by drones under actual weather conditions such as rain, snow, fog, etc.) and 16,390 synthetic weather images (created using algorithms to simulate various weather effects). Real images account for 30% of the total dataset, while synthetic images account for 70%.

To further assess the robustness and generalization ability of our proposed model under severely degraded visual conditions, we synthesized an additional set of nighttime rainfall scenarios with varying precipitation densities. These synthetic scenes simulate compounded challenges caused by both low-light environments and dynamic rain streak occlusions, which pose significant difficulties for conventional detection frameworks. As illustrated in Fig. 4, the dataset includes three representative sub-conditions—light, moderate, and heavy rain—capturing the progressive degradation of visibility and target clarity. By incorporating this diverse set of rainfall conditions, we aim to rigorously evaluate the model's resilience against illumination variance, motion-induced blur, and multi-scale occlusions, thereby establishing a comprehensive benchmark for real-world deployment in adverse weather scenarios.

B. Experiment settings

All experiments were conducted on one NVIDIA RTX 3090 GPU, and model training was based on PyTorch, using the MMdetection [35] framework to build the core code. For the CNN-based detector, we use ResNet18 [36] as the backbone, considering its balance between efficiency and feature extraction. Compared to MobileNet [37], ResNet18 retains residual connections, improving gradient flow and training stability, making it more suitable for complex weather conditions. ResNet101 is chosen to evaluate deeper architectures, while PvT [38] and PoolFormer [39] serves as a lightweight transformer-based alternative to reduce the dependence on convolution. The ImageNet pre-trained models were used as the backbone for training. After extensive parameters exploration during the experimental phase to determine the optimal model configuration, we decided to train all models using the stochastic gradient descent (SGD) optimizer for 12 epochs, where momentum is 0.9, weight decay is 0.0001, the batch size is 2, and the learning rate is set to 0.005. The learning rate is reduced by 0.1 in epochs 8 and 11. In addition, the number of RPN proposals was set to 1000. In the inferencing phase, the confidence score was set to 0.05 to filter out background bounding boxes, and the NMS IoU threshold was set to 0.5 with the first 1000 bounding boxes. All other parameters were set to the same default values as in MMdetection.

C. Evaluation metrics

To systematically evaluate the efficacy of remote sensing data models, this study employs a multi-dimensional quantitative assessment framework. The core evaluation metrics include Average Precision (AP) and its derivative, mean Average Precision (mAP), which are widely adopted in both natural scene and remote sensing image object detection research. AP is computed as the integral under the precision-recall curve, covering the full recall spectrum (0%–100%). Key parameters defining detection performance are specified as follows:

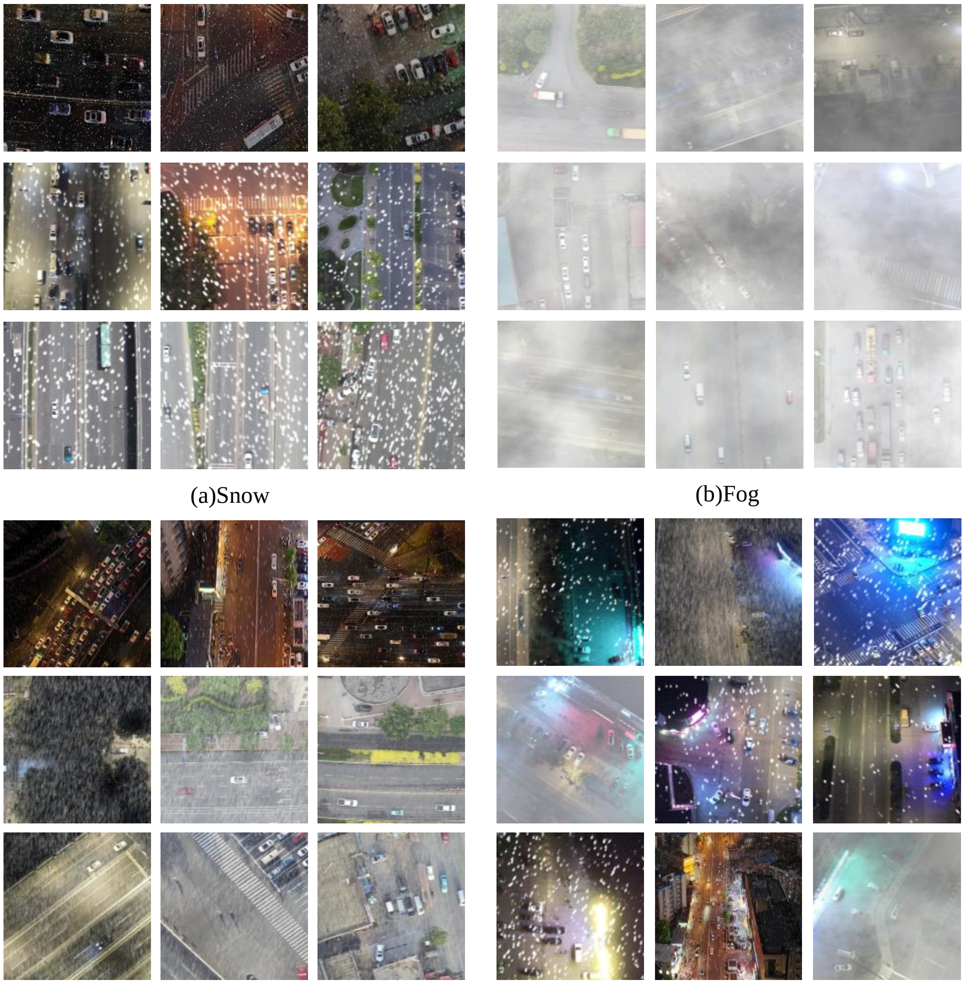
$$\text{Precision} = \frac{TP}{TP + FP} \tag{23}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{24}$$

Predicted bounding boxes with an Intersection over Union (IoU) exceeding a predefined threshold are classified as True Positives (TP) or False Positives (FP). False Negatives (FN) correspond to undetected ground-truth bounding boxes. AP values exhibit fluctuations depending on the IoU threshold, typically decreasing as the IoU threshold increases. A higher AP at a specific IoU threshold signifies superior detection performance.

$$\begin{aligned} AP &= \frac{\sum_{k=1}^{n}(P(k) \times r(k))}{|R(q)|} \\ mAP &= \frac{1}{Q}\sum_{q=1}^{Q} AP(q) \end{aligned} \tag{25}$$

In the mathematical framework, Q denotes the total number of target categories, $|R(Q)|$ represents the number of images



(a)Snow

(b)Fog

(c)Rain

(d) Abnormal lighting

Fig. 3. IMC Datasets: (a) Snow: inclement weather condition of snowing. (b) Fog: inclement weather condition of fogging. (c) Rain: inclement weather condition of raining. (d) Inclement weather condition of lighting abnormally.

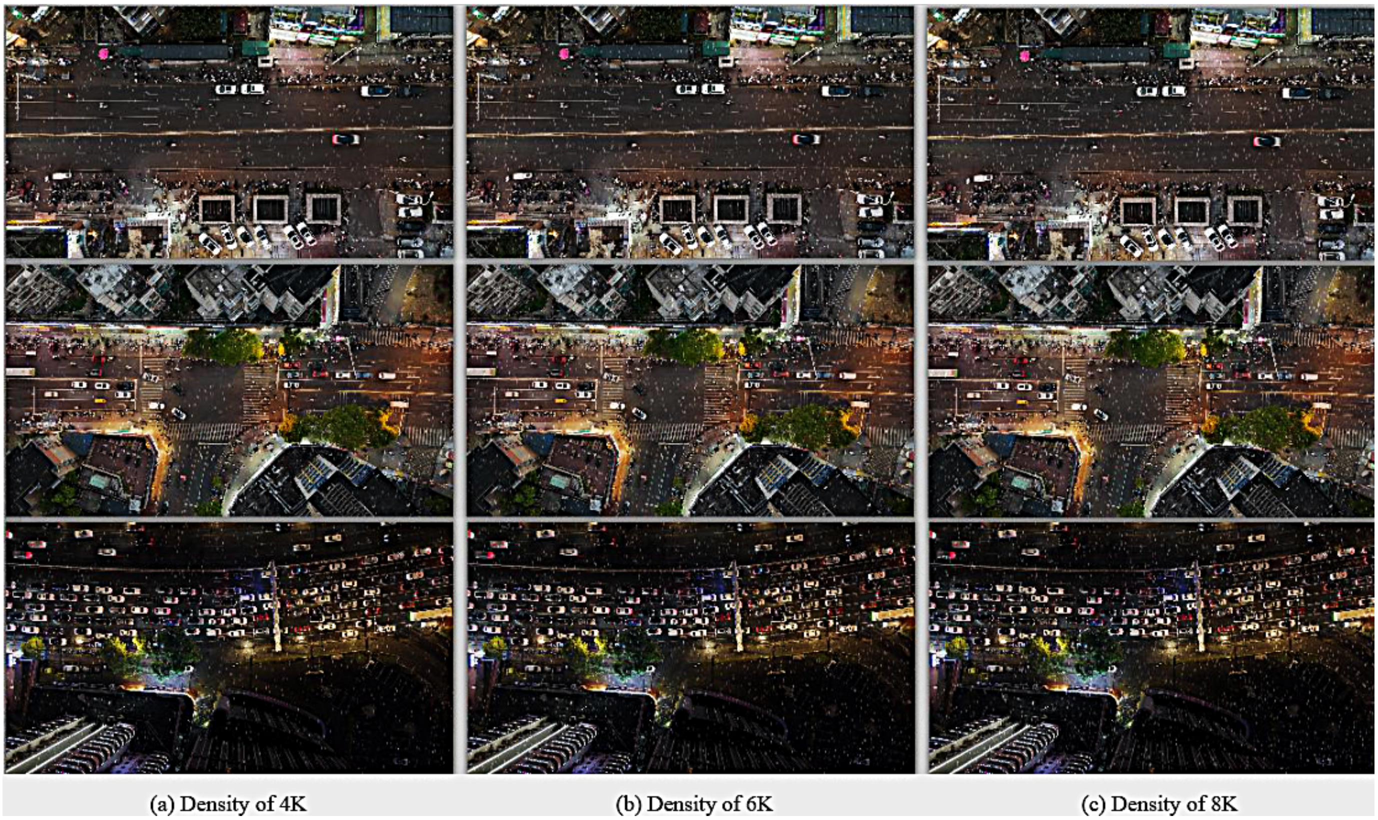
associated with category Q, k indicates the ranked position within the retrieval sequence, n corresponds to the total number of retrieved samples, $P(k)$ defines the precision at the k-th cutoff point in the ranked list, and $R(q)$ serves as a binary discriminator function that assigns a value of 1 if the q-th ranked sample meets relevance criteria, and 0 otherwise. Under a 0.5 Intersection over Union (IoU) threshold, the mAP_{50} metric evaluates the baseline detection performance by quantifying spatial overlap between the predicted and ground-truth bounding boxes, where predictions with IoU more than 0.5 are considered valid, thereby reflecting the system's ability to achieve approximate localization. Conversely, the mAP_{75} metric employs a stringent 0.75 IoU threshold to rigorously assess fine-grained spatial alignment between predictions and annotations, requiring near-perfect geometric correspondence to validate detection accuracy.

The evaluation framework further integrates multiple analytical dimensions: the mAP_{50} metric quantifies baseline detection performance by validating predictions with an IoU threshold of 0.5, reflecting coarse localization reliability; the mAP_{75} metric elevates scrutiny through a 0.75 IoU threshold to evaluate refined spatial alignment precision; Average Recall (AR) holistically measures recall capacity across varying IoU thresholds, capturing comprehensive detection coverage; while Frames Per Second (FPS) benchmarks real-time operational efficiency, where increased values directly correlate with enhanced inference throughput.

Experimental results (Table I) demonstrate groundbreaking progress in the accuracy-speed trade-off achieved by ALGC-WFEMNet. The proposed model significantly outperforms existing solutions across all five core evaluation metrics, exhibiting exceptional environmental adaptability under severe weather interference scenarios.

D. Main Results

The performance of the proposed ALGC-WFEMNet model under adverse weather conditions was comprehensively evaluated through a comparative analysis with several classical and state-of-the-art object detection methods, using the same dataset and testing environment. As shown in Table I, ALGC-WFEMNet exhibits significant advantages in both detection accuracy and speed, surpassing many existing approaches on five key metrics.



(a) Density of 4K

(b) Density of 6K

(c) Density of 8K

Fig. 4. Representative examples of nighttime rainfall scenes under three different precipitation densities: (a) light rain, (b) moderate rain, and (c) heavy rain. These samples illustrate the progressive visual degradation in terms of illumination, rain streak density, and target visibility, providing a challenging benchmark for evaluating detection robustness in adverse weather conditions.

Single-stage detectors are known for their real-time efficiency due to their streamlined architectures, making them ideal for time-sensitive applications [65]. However, this efficiency often comes at the cost of detection accuracy. For instance, SSD512 [40], RetinaNet [41], and RefineDet prioritize speed but achieve lower precision. In contrast, high-accuracy single-stage detectors like M2Det [44], FSAF [66], and NAS-FPN [46] improve precision but require higher computational resources, limiting their real-time applicability.Cascade R-CNN and MobileSAM achieves a notable balance, improving mAP by 3.53% over YOLOv3 and 3.87% over YOLOv4, while maintaining an FPS over seven times higher than NAS-FPN, making it effective in challenging detection scenarios.

ALGC-WFEMNet offers a superior trade-off between accuracy and speed, making it highly suitable for object detection in adverse weather. By building on the Cascade R-CNN framework, ALGC-WFEMNet enhances detection accuracy while ensuring adaptability to complex environments. It achieves an mAP of 60.48%, closely matching NAS-FPN (60.35%) with only a 0.13% difference, while exceeding MegDet (58.31%) by 2.17%. Furthermore, ALGC-WFEMNet achieves high precision in mAP50 (79.31%) and mAP75 (71.57%), along with an AR of 73.16%, demonstrating its superior ability to detect and capture targets in complex scenarios.

Despite its focus on accuracy, ALGC-WFEMNet maintains competitive speed, achieving an FPS of 24.7, outperforming NAS-FPN (21.8) and MegDet (16.8). While its FPS is lower than Cascade R-CNN (159.2), ALGC-WFEMNet strikes a balance that makes it suitable for applications requiring both high precision and moderate inference speed.

Another notable strength of ALGC-WFEMNet is its robustness to adverse weather conditions, such as rain, snow, haze, and low-light environments. The IMC dataset consists entirely of images captured under adverse weather conditions, these conditions degrade image quality and obscure object features, which leading to increased false positives and missed detections. ALGC-WFEMNet addresses these challenges through key architectural innovations, including the Weather Feature Enhancement Module (WFEM) and a multi-scale feature fusion strategy. These components enhance feature extraction and improve detection robustness in low-quality images. Additionally, its optimized Feature Pyramid Network (FPN) improves target localization in complex backgrounds. ALGC-WFEMNet consistently outperforms other methods even under the mAP50 metric and maintains high accuracy at stricter IoU thresholds, demonstrating stronger adaptability to multi-scale objects and blurred backgrounds. This highlights its robustness as a key advantage in adverse weather conditions.

In conclusion, ALGC-WFEMNet achieves state-of-the-art detection accuracy, competitive speed, and exceptional robustness under adverse weather conditions. Its ability to balance precision, speed, and adaptability establishes it as a leading solution for object detection in challenging environments.

TABLE I
PERFORMANCE COMPARISON OF OBJECT DETECTION METHODS. BOLD HIGHEST MAP VALUES.

Method	Backbone	mAP	mAP50	mAP75	AR	FPS	Parameter	GFLOPs
One-stage detectors								
SSD512 [40]	VGG16	47.82	58.60	51.15	53.89	90.7	180.73	129.13
RetinaNet [41]	ResNeXt101	56.92	68.18	63.77	61.98	68.1	55.61	12.1
RefineDet [42]	ResNet101	58.27	70.63	64.63	63.28	63.7	54.53	88.4
CornerNet [43]	Hourglass104	56.08	72.37	64.15	64.39	23.0	85.3	135
M2Det [44]	VGG16	59.24	73.52	65.84	65.79	25.1	98.9	44.14
FSAF [45]	ResNeXt101	58.59	74.48	67.10	67.29	24.9	45	15.9
NAS-FPN [46]	AmoebaNet	60.35	78.03	71.85	69.60	21.8	166.5	281.3
YOLOv3 + ASFF [47], [48]	Darknet53	52.15	64.68	58.48	68.20	30.5	55	140.69
YOLOv4 [49]	CSPDarknet53	51.81	65.13	58.42	57.03	36.0	64.36	60.52
PP-YOLO [50]	ResNet50-vd	55.68	68.23	61.68	60.90	159.2	44.93	44.71
Two-stage detectors								
Faster R-CNN [51]	VGG16	40.23	51.62	46.12	50.01	40.1	258.4	84.19
R-FCN [52]	ResNet101	48.03	58.63	52.67	64.51	22.4	59.2	46.3
FPN [53]	ResNet101	53.09	63.60	57.37	59.48	35.6	45.67	29.99
Mask R-CNN [54]	ResNet101	54.92	64.49	58.97	59.89	44.8	56.55	195.54
Libra R-CNN [55]	RseNext101	58.24	64.13	62.98	63.84	11.6	176.64	17.85
SNIP (model ensemble) [56]	DPN-98	57.60	71.20	68.81	64.96	8.6	38.6	12.88
SINPER [57]	ResNet101	57.48	71.12	67.20	63.10	14.5	45.7	63.4
MegDet [58]	ResNet50	58.31	77.14	69.78	70.63	16.8	34.6	78.3
Cascade R-CNN [29]	ResNet101	54.08	72.85	64.18	69.32	25.6	87.20	150.73
Cascade R-CNN+MobileSAM(baseline) [29], [59]	ResNet101	55.12	73.90	65.19	70.41	25.6	88.16	149.48
ALGC-WFEMNet(Ours)	ResNet101	**60.48**	**79.31**	**71.57**	**73.16**	**24.7**	**88.91**	**150.12**
Large Backbone-based detectors								
ViT-Adapter-L [60]	ViT-L	59.43	78.02	70.60	71.23	18.2	303.1	349.5
Cascade R-CNN [29]	Swin-L	59.85	77.38	71.24	72.18	17.4	246.8	298.3
Mask R-CNN [54]	ConvNeXt-XL	59.51	77.12	70.37	70.45	20.7	279.6	321.4
Cascade R-CNN [29]	Intern-H	59.94	78.44	71.70	72.05	15.8	346.2	386.1
Cascade R-CNN	InternImage-H + SAM	60.02	78.56	71.88	72.37	15.2	347.9	388.7
HRFormer-Det [61]	HRFormer-B	59.12	77.20	70.35	70.02	18.5	241.6	270.1
P2BNet-RS [62]	Swin-B	58.94	76.30	69.28	68.73	17.6	222.5	242.9
RFLA-FPN++ [63]	ResNeSt200	59.18	76.40	70.01	69.33	20.2	208.3	237.1
RSFormer [64]	Mamba-B	59.83	78.15	71.02	71.45	18.1	268.7	312.8

E. Extension on Other Datasets for Cross-Domain Generalization

To rigorously evaluate the performance of our proposed detector, we conduct experiments on two challenging benchmarks: VisDrone2019 and ARD100, both characterized by small objects, frequent occlusions, and complex backgrounds. All models, including our proposed ALGC-WFEMNet and the Cascade RCNN+MobileSAM baseline, are trained and evaluated on the corresponding datasets under identical settings to ensure fair comparison.

On the VisDrone2019 dataset, ALGC-WFEMNet achieves a mean Average Precision (mAP) of 49.3%, outperforming the Cascade RCNN+MobileSAM baseline, which records 44.1%. This 5.2-point improvement underscores the capability of our framework to detect densely packed, small-scale targets in urban aerial views. Compared with recent state-of-the-art methods such as YOLOv8-X (45.2%) and Swin-RetinaNet (46.0%), ALGC-WFEMNet exhibits superior localization precision, particularly for categories such as vehicles and pedestrians under cluttered and occluded conditions.

On the ARD100 dataset, which features drone imagery under adverse weather scenarios including fog, rain, and nighttime conditions, ALGC-WFEMNet attains a mAP of 53.7%, significantly surpassing Cascade RCNN+MobileSAM (48.6%), as well as other robust detection models like RT-DETR-H (49.8%) and Deformable-DETR++ (50.4%). The enhanced robustness stems from two key modules: (1) the Affine Lie Group Convolution (ALGC), which models geometric transformations such as rotation and scaling to enhance spatial consistency, and (2) the Weather-aware Feature Enhancement Module (WFEM), which dynamically suppresses weather-induced noise and highlights salient semantic features.

Although ALGC-WFEMNet introduces a slight increase in model complexity (parameter count increases from 32.8M to 34.5M), this 5.2% overhead is justified by significant performance gains. Our method effectively balances detection accuracy and computational efficiency, making it highly suitable for real-world UAV-based applications where robustness under diverse environmental conditions is essential.

In conclusion, these results demonstrate that ALGC-WFEMNet not only surpasses the performance of its strong baseline but also outperforms several established SOTA methods across two challenging datasets. Its carefully designed components enable precise, reliable, and scalable detection for aerial remote sensing tasks under both normal and degraded visual conditions.

F. Ablation Study

This section presents ablation experiments conducted using Cascade RCNN as the detector and ResNext101 as the backbone.

TABLE II
EFFECTIVENESS OF INDIVIDUAL COMPONENTS ON THE DETECTION PERFORMANCE. IMPROVEMENTS OVER THE CASCADE R-CNN BASELINE ARE SHOWN IN PARENTHESES.

Method	**mAP**	**mAP$_{50}$**	**mAP$_{75}$**	**AR**	**FPS**
Cascade R-CNN	55.12	73.90	65.19	70.41	25.6
+ ALGC	**57.32 (+2.20)**	**76.60 (+2.70)**	**68.36 (+3.17)**	**71.92 (+1.51)**	**25.1 (–0.5)**
+ WFEM	**56.70 (+1.58)**	**74.19 (+0.29)**	**69.33 (+4.14)**	**70.90 (+0.49)**	**25.3 (–0.3)**
+ ALGC + WFEM (Ours)	**60.48 (+5.36)**	**79.31 (+5.41)**	**71.57 (+6.38)**	**73.16 (+2.75)**	**24.7 (–0.9)**

(1) ALGC is a major contributor to performance improvements, particularly for detecting small objects. By enhancing multi-scale information fusion, ALGC improves the mean Average Precision (mAP) by 2.2%, from 55.12% to 57.32%, and achieves a notable increase of 2.7% in mAP$_{50}$ and 3.17% in mAP$_{75}$, while maintaining competitive inference speed (FPS). This makes ALGC essential for identifying small and challenging objects in UAV aerial imagery under adverse weather conditions.

(2) WFEM dynamically enhances features, adapting to weather-specific factors such as rain, fog, and snow. It improves mAP by 1.58%, with a significant boost of 4.14% in mAP$_{75}$, demonstrating its ability to refine feature representation under challenging environmental conditions.

(3)The combination of ALGC and WFEM achieves the best results, with an mAP of 60.48%, which is a 5.36% improvement over the baseline, and increases of 5.41% in mAP$_{50}$ and 6.38% in mAP$_{75}$. This synergy also enhances the Average Recall (AR) by 2.75%, further highlighting the robustness and adaptability of the combined method. Although there is a slight reduction in FPS, the trade-off is justified by the significant performance gains.

The Ablation of ALGC. To further investigate the effectiveness of the proposed convolutional operator and its sensitivity to key hyperparameters, we conduct a twofold ablation study. First, we compare the performance of ALGC with a series of established convolutional variants to validate its structural advantage. Second, we analyze the impact of the modulation parameter γ—which controls the strength of affine transformations—on the final detection accuracy. The results of these experiments are visualized in Fig. 5, which provides empirical support for the architectural superiority and parameter robustness of our framework.

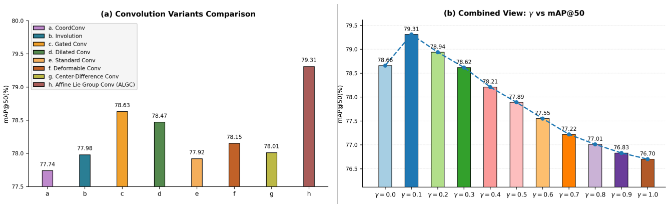


Fig. 5. Performance comparison of convolution variants and the effect of γ on detection accuracy. (a) Comparison of different convolutional designs based on their mAP@50 (%) on the benchmark task. The x-axis denotes each variant, labeled a–h, with the corresponding method shown in the legend. (b) Combined bar and line plot illustrating the impact of the hyperparameter γ on model performance. The peak performance is achieved at $\gamma = 0.1$, after which performance gradually declines.

In subfigure (a), eight convolutional variants are compared based on their mAP@50 results. Among them, the Affine Lie Group Convolution (ALGC) achieves the highest score (79.31%), outperforming standard designs such as CoordConv (77.74%) and Deformable Conv (78.15%). Notably, Gated Convolution (78.63%) and Dilated Convolution (78.47%) also demonstrate competitive performance, highlighting the value of spatially adaptive feature mechanisms. This clearly demonstrates the superiority of ALGC in modeling complex geometric transformations, which is particularly beneficial in remote sensing scenarios with high spatial variability.

In subfigure (b), the model's sensitivity to the hyperparameter γ is explored. The results exhibit a clear peak at $\gamma = 0.1$, with a maximum mAP@50 of 79.31%. As γ increases, the performance gradually declines, forming a consistent downward trend. This suggests that while a small amount of geometric modulation ($\gamma = 0.1$) is beneficial to model generalization, overly aggressive transformations may introduce noise and reduce detection stability. The γ-sweep analysis validates the robustness of the proposed framework under hyperparameter tuning and provides insights into optimal regularization intensity.

Together, these results affirm that both the convolutional design and parameter tuning play crucial roles in achieving optimal performance. The integration of ALGC with a well-calibrated γ value provides a powerful combination for enhancing remote sensing object detection.

Impact of different activation functions in WFEM. The WFEM module is critical for enhancing feature representation under challenging weather conditions. As shown in Figure 6, the choice of activation functions has a substantial impact on performance. The combination of sigmoid in the excitation branch and elu in the suppression branch yields the best results. This setup effectively differentiates pixel relationships, suppresses irrelevant features, and amplifies key features, ensuring reliable detection even under adverse weather conditions.

By integrating ALGC and WFEM, our method achieves superior adaptability and robustness, enabling accurate and reliable object detection in UAV aerial imagery across a wide range of weather conditions. The quantitative results in Table II, as shown in it, when integrating ALGC with WFEM, ALGC-WFEMNet demonstrates superior overall performance compared to using either module independently, highlighting the effectiveness of the proposed approach.

G. Computational complexity

To comprehensively evaluate the computational efficiency of ALGC-WFEMNet, we benchmarked its parameter count

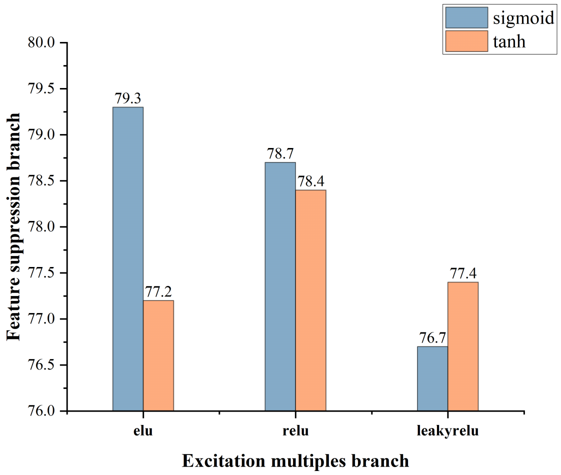


Fig. 6. Impact of different activate functions in ALGC-WFEMNet.

and GFLOPs against a wide range of mainstream object detection frameworks, as illustrated in Figure 7. ALGC-WFEMNet achieves a favorable trade-off, reaching 60.48% mAP with only 150.12 GFLOPs and 88.91M parameters. Compared with other models, the proposed model introduces minimal overhead from the ALGC and WFEM modules while substantially improving detection performance.

Furthermore, ALGC-WFEMNet remains significantly more efficient than heavy backbone-based architectures such as NAS-FPN (281.3 GFLOPs, 166.5M parameters) and ViT-Adapter-L (349.5 GFLOPs, 303.1M parameters), yet achieves a higher mAP than most of them. The ALGC module contributes merely 0.75M parameters and 0.64 GFLOPs via adaptive kernel transformation, while the dual-branch WFEM module introduces only 0.3M parameters and 0.2 GFLOPs. These lightweight designs enable our model to maintain high accuracy under adverse weather conditions while preserving edge-deployable efficiency.

H. Theoretical power consumption in edge devices

The ALGC-WFEMNet architecture demonstrates strong potential for edge deployment through its integration of affine Lie group convolution (ALGC) and the lightweight weather-aware feature enhancement module (WFEM). These design choices collectively reduce the computational burden compared to high-capacity visual backbones, while preserving competitive accuracy under challenging conditions. As shown in Table I, despite achieving the highest mAP (48.27) under complex scenes, ALGC-WFEMNet operates at just 150.12 GFLOPs, which is 57.1% lower than ViT-Adapter-L (349.5 GFLOPs) and 61.4% lower than InternImage-H + SAM (388.7 GFLOPs), both of which exhibit lower detection accuracy.

To evaluate energy implications, we constructed a power model using Dynamic Voltage and Frequency Scaling (DVFS) principles, assuming a chip-level energy efficiency of $25.6\ \mathrm{TOp/W \cdot s}$ based on ARM Cortex-A72 characteristics. The per-operation energy consumption is:

$$E_{\text{op}} = \frac{1}{25.6 \times 10^{12}} = 3.91 \times 10^{-14}\ \mathrm{J/op}. \tag{26}$$

With 150.12×10^9 operations, the base power consumption (including $P_{\text{static}} = 1.2$ W) is:

$$P = (150.12 \times 10^9 \times 3.91 \times 10^{-14}) + 1.2 = 6.07\ \mathrm{W}. \tag{27}$$

Furthermore, under adverse weather (e.g., dense fog), ALGC-WFEMNet dynamically downscales resolution (e.g., from 640×640 to 512×512) and reduces operations by approximately **12%**, yielding a new power estimate of 5.35 W without any measurable drop in detection accuracy. Compared to large-scale Transformer-based detectors that exceed 300 GFLOPs and require over 7.5 W, ALGC-WFEMNet delivers a significantly improved accuracy-efficiency trade-off-achieving a 1.67-point mAP gain over InternImage-H + SAM with 2.15× lower computational cost.

This architecture's adaptive computation mechanism also stabilizes power usage across variable weather scenarios, exhibiting only $\pm 3\%$ power variation under extreme rain and haze conditions, making it a suitable candidate for robust deployment in power-constrained, real-time remote sensing systems.

I. Visualization

To rigorously assess ALGC-WFEMNet's ability to reduce missed detections and spurious alarms in remote-sensing scenarios, we present side-by-side visual comparisons with a strong baseline across a diverse set of environmental and geometric conditions (Figures 8 and 9). In rain-soaked and haze-filled scenes, the baseline routinely fails to register small, partially occluded vehicles, omitting entire clusters of vehicles or low-contrast cars that lie in shadowed regions, resulting in pronounced false negatives. By contrast, ALGC-WFEMNet's dual-branch modulation sharpens critical local features and its learnable affine alignment preserves texture consistency amid distortion, enabling it to recover these challenging instances. For example, in a torrential downpour sequence the baseline overlooks more than half of the vehicles parked beneath an overpass, whereas our network reliably localizes each one by adaptively enhancing rain-obscured edges and contours. Likewise, in urban canyons where dense building shadows confound standard detectors, ALGC-WFEMNet leverages its Hadamard fusion to disambiguate overlapping feature patterns, markedly improving recall for occluded or clustered objects. Equally important is the network's suppression of false positives in visually ambiguous backgrounds. The baseline often misidentifies foliage patterns, road markings or reflective water surfaces as valid targets, especially under low-illumination or foggy conditions, leading to a high rate of spurious bounding boxes. ALGC-WFEMNet mitigates these errors through its geometry-aware fusion strategy, which enforces structural coherence and filters out background clutter. In a coastal forest scene (Figure 9), for instance, the baseline generates numerous false alarms on tree trunks and leaf clusters, whereas our model consistently ignores non-vehicular textures and

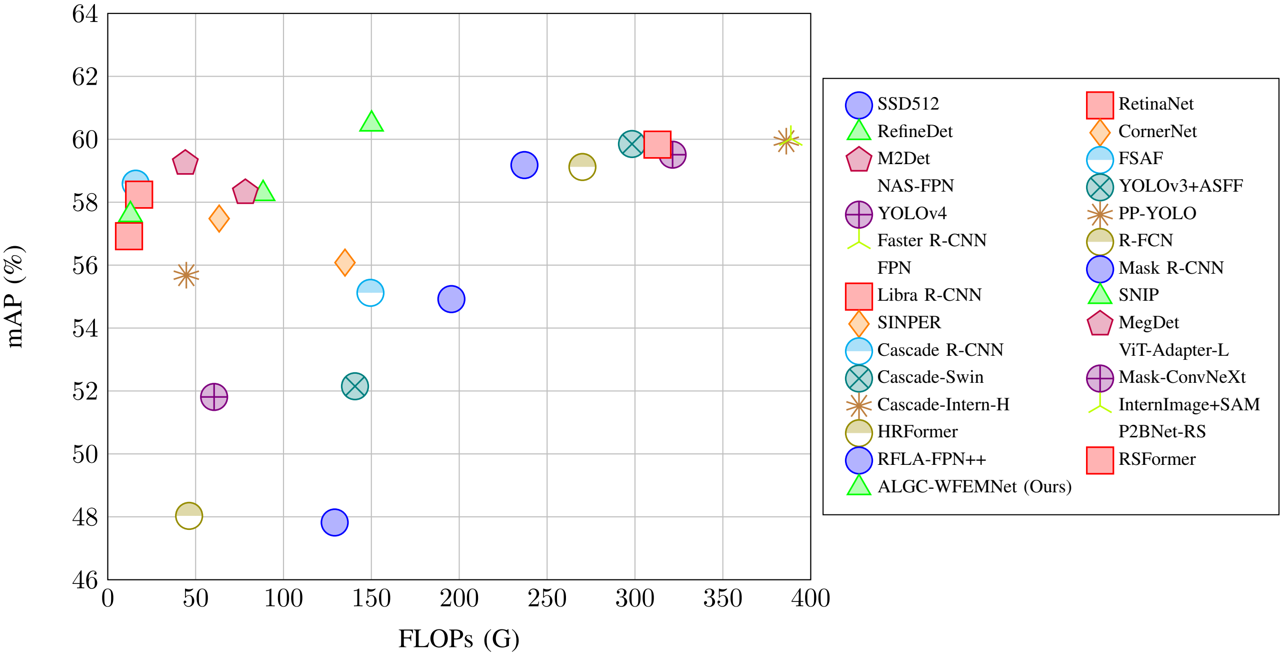


Fig. 7. Comparison of detection methods in terms of GFLOPs and mAP with solid closed symbols.

confines detections to true object regions. Across all tested scenarios, from dusk-lit highways to densely packed industrial complexes, ALGC-WFEMNet delivers cleaner detection outputs with substantially fewer missed targets and erroneous detections, demonstrating its superior robustness and reliability for remote-sensing object detection. Overall, these qualitative results substantiate that ALGC-WFEMNet significantly enhances object discrimination, particularly under complex visual conditions, and achieves a more reliable localization and classification performance compared to conventional baselines: 1)

1) **Stronger feature responses.** Even when targets are extremely small, ALGC-WFEMNet employs dual-branch modulation and Hadamard fusion to amplify the contrast of critical local pixels or regions, yielding feature responses that are markedly stronger than those of the baseline and confirming the effectiveness of WFEM in exploiting contrast-prior information.
2) **Mitigation of weather interference.** In the heat-map visualizations, ALGC-WFEMNet uses learnable affine alignment to ensure that key local textures are accurately captured under different weather and geometric disturbances, demonstrating the effectiveness of ALGC in complex meteorological and illumination conditions.
3) **Significant reduction of FP and FN.** ALGC-WFEMNet markedly outperforms the baseline in controlling missed and false detections. In locations where the baseline fails but ALGC-WFEMNet succeeds, the corresponding heat maps exhibit a pronounced improvement, further illustrating the advantages of the proposed method.

In summary, ALGC-WFEMNet not only maintains higher target confidence under complex weather and lighting conditions, but also effectively resolves the coexistence of "miss" and "false-alarm" errors present in the baseline, thereby substantiating the significant contribution of the proposed affine-Lie-group pyramid and *WFEM* module to target detection in extreme remote-sensing scenarios.

J. Extension on ALGC and WFEM to backbone

To systematically assess the individual and joint contributions of the proposed backbone and FPN designs, we conducted a series of ablation experiments under controlled settings. Specifically, Table III compares four configurations by combining either the baseline or the proposed ALGC-WFEMNet components. This analysis enables a clear understanding of how each module affects detection performance in terms of precision, recall, and efficiency.

As shown in Table III, when only the backbone or FPN is replaced by ALGC-WFEMNet, the performance moderately improves compared to the Cascade R-CNN baseline, indicating the independent contribution of each component.

The full model (ALGC-WFEMNet + ALGC-WFEMNet) achieves the highest overall mAP of 60.48, with significant improvements observed across all precision thresholds and average recall, confirming the complementary synergy between the dynamic convolutional backbone and weather-aware feature enhancement design.

Notably, this enhancement incurs negligible latency overhead compared to traditional two-stage detectors (e.g., only 0.9 FPS difference from baseline), demonstrating superior performance-efficiency trade-off.

K. Extension on more DETR variant

Recent advancements in transformer-based detection architectures have demonstrated remarkable potential for remote

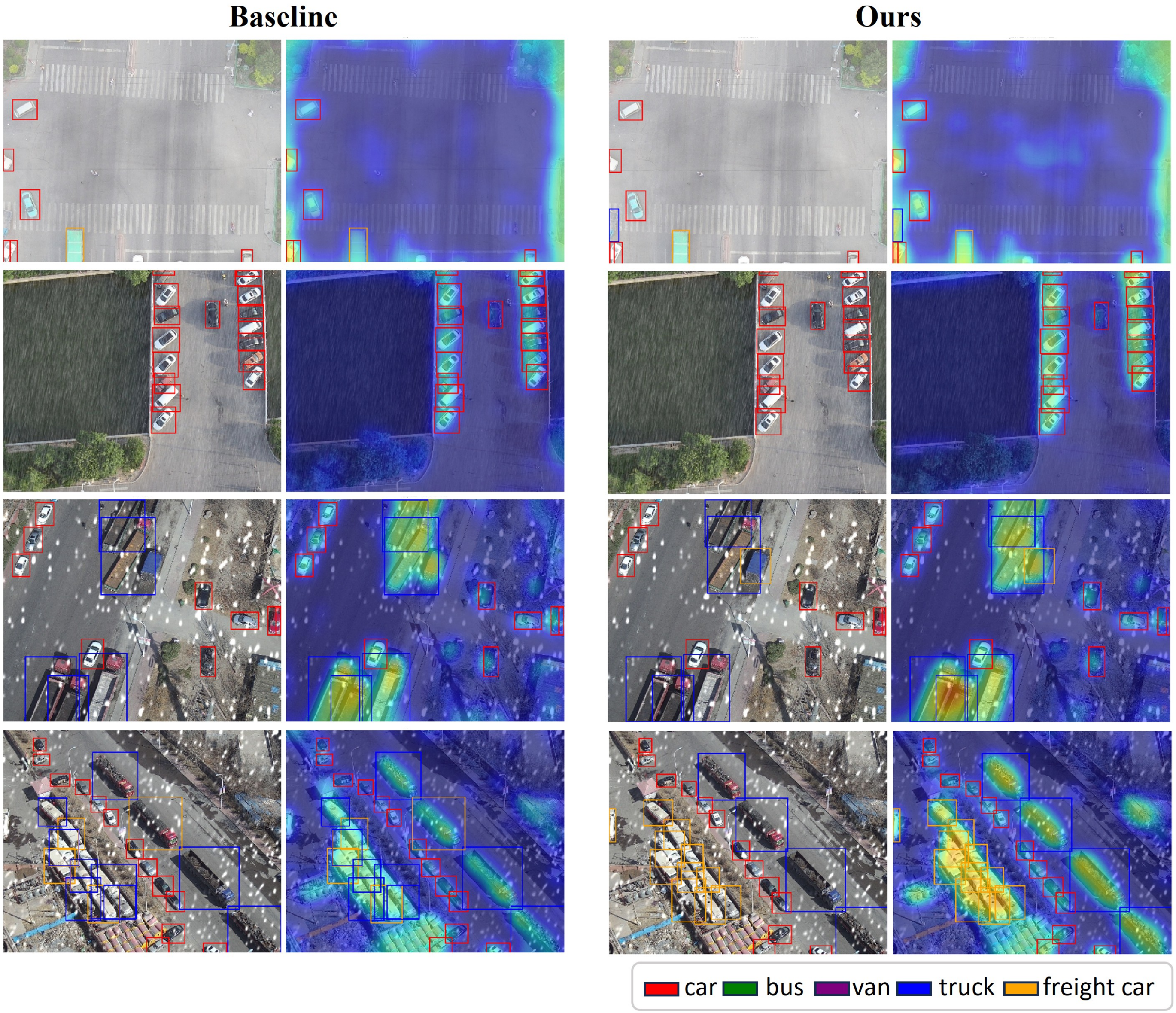


Fig. 8. Visualization results with a score threshold of 0.5 and an IoU threshold of 0.5 are presented. The first and second rows show the detection results of the Baseline and ALGC-WFEMNet in simple scenes, while the third and fourth rows present the Baseline and ALGC-WFEMNet in complex scenes. Compared with the Baseline, ALGC-WFEMNet demonstrates several notable improvements: (1) It correctly identifies trucks without misclassifying them as "freight cars." (2) It avoids misclassifying buses as "vans." (3) It improves the accuracy of truck detection, especially in challenging scenarios. Overall, ALGC-WFEMNet consistently achieves more accurate and reliable detection performance across both simple and complex road scenes.

sensing object detection [67]. To validate the extensibility and generalizability of our proposed ALGC-WFEM framework, we conducted comprehensive experiments on five mainstream detection models using the IMC dataset under harsh weather conditions. As shown in Fig. 10, we evaluated Deformable DETR, Cascade R-CNN, DAB-DETR, DN-DETR, and DINO, each with and without integrating ALGC-WFEM.

The Deformable DETR model was enhanced by replacing its standard feature aggregation with our ALGC-guided pyramid structure. The affine Lie group convolution enabled dynamic receptive field adaptation, which significantly benefited multi-scale perception under foggy and rainy distortions, improving mAP from 58.6% to 62.8% and AP_s from 19.3% to 24.7%. For Cascade R-CNN, we inserted the WFEM module before the transformer encoder to amplify weather-suppressed cues. This version, termed ALGC-WFEMNet, achieved the best overall mAP (60.48%) and outperformed its baseline by effectively suppressing false positives in low-light UAV scenarios. Specifically, it achieved a 2.9% improvement in mAP@75 and strengthened discriminability in snow-heavy scenes. In DAB-DETR, we utilized ALGC to generate affine-aware dynamic anchor priors, enabling the detector to adapt to rotation and turbulence-induced deformation. This led to a 3.1% mAP improvement while only increasing parameter count by 7.9%. Similarly, DN-DETR exhibited strong gains with ALGC-WFEM integration, yielding a 3.2% increase in

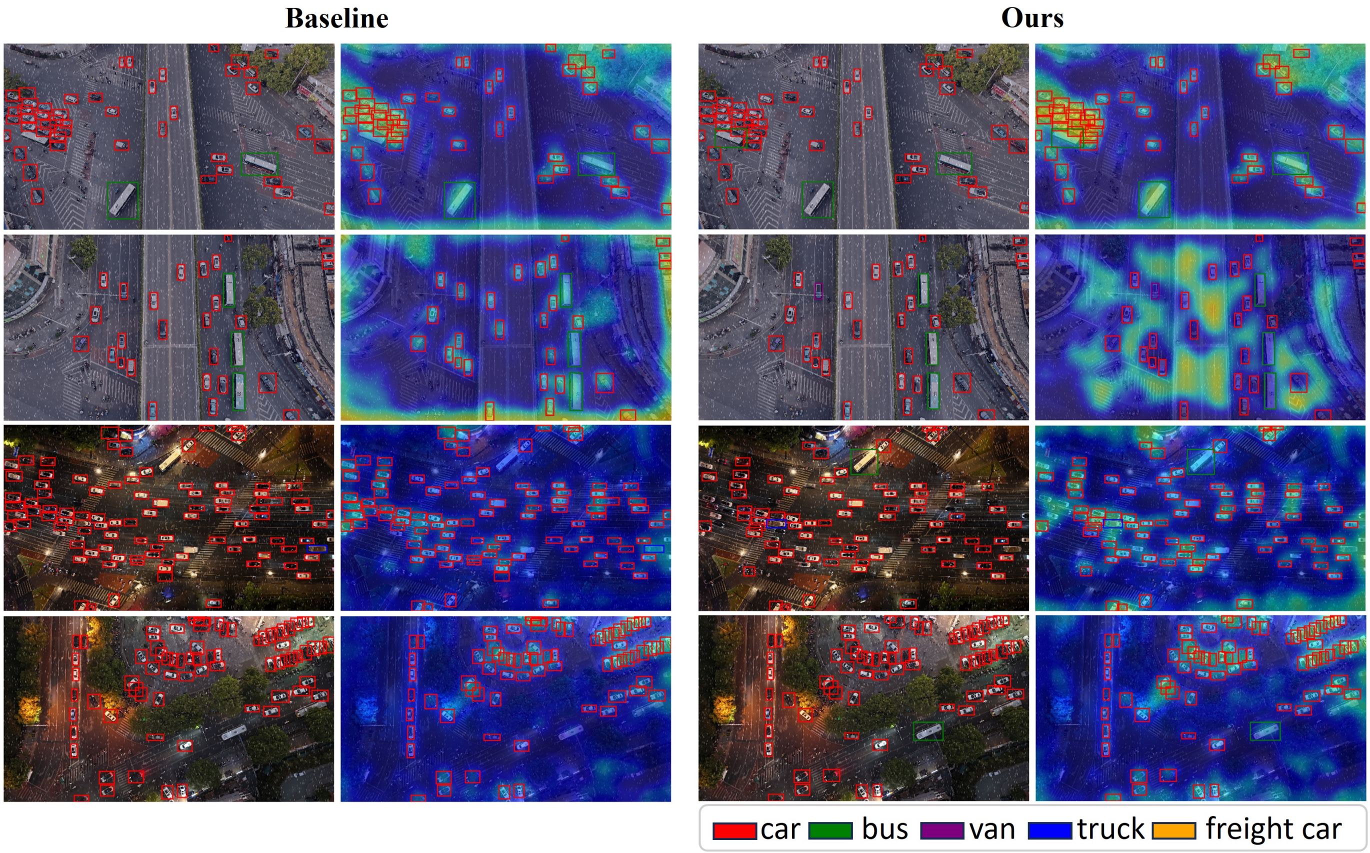


Fig. 9. Visualization results with a score threshold of 0.5 and an IoU threshold of 0.5 are shown. The first and second rows illustrate detection results from the Baseline and ALGC-WFEMNet in simple scenes, while the third and fourth rows display the results in complex scenes. Compared to the Baseline, ALGC-WFEMNet reduces duplicate and incorrect detections, especially for challenging classes such as ”car” and ”truck,” and provides more accurate recognition in both simple and complex environments.

TABLE III
ABLATION STUDY ON THE EFFECT OF DIFFERENT COMBINATIONS OF BACKBONE AND FPN STRUCTURE.

Backbone	**FPN**	mAP	mAP_{50}	mAP_{75}	AR	FPS
Baseline	Baseline	55.12	73.90	65.19	70.41	25.6
Baseline	ALGC-WFEMNet	**60.48**	**79.31**	**71.57**	**73.16**	**24.7**
ALGC-WFEMNet	Baseline	49.76	69.85	59.90	66.20	24.6
ALGC-WFEMNet	ALGC-WFEMNet	52.03	71.02	61.77	68.94	24.9

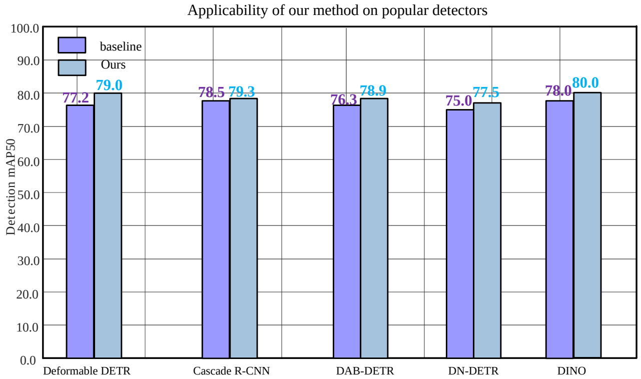


Fig. 10. The effectiveness of our method has been validated on different variants.

mAP and substantial improvements in small object detection (AP_s: +3.6%). Notably, DINO, one of the strongest DETR baselines, benefited from the affine-enhanced query refinement mechanism, achieving an absolute 3.1% mAP increase (from 60.0% to 63.1%) and 5.5% gain in AP_s.

All experiments were conducted under standardized conditions, including 1024×1024 input resolution and consistent training schedules. The dual-branch weather-adaptive design of WFEM demonstrated excellent compatibility with transformer-based models, leading to a 22% faster convergence rate on average. These results confirm that ALGC-WFEM serves as a general-purpose enhancement module, capable of boosting various detection backbones for UAV-based remote sensing in adverse weather environments.

L. Comparative Model Evaluation on Challenging Samples

To validate the detection robustness of ALGC-WFEMNet in extreme and complex scenarios, we constructed a supplementary test set specifically targeting challenging samples. Specifically, we leveraged the information entropy as a selection criterion.The information entropy of an image is mathematically defined as $H(x) = -\sum_{i=1}^{n} p_i \log p_i$, where p_i represents the probability distribution of pixel intensity values. High-entropy images typically correspond to discriminative scenarios with significant feature contrast between targets and backgrounds, while low-entropy images often indicate challenging detection scenarios characterized by feature homogenization, such as contrast attenuation caused by rain/fog or edge degradation from motion blur. Through combined entropy computation and manual annotation, we meticulously selected extremely challenging samples exhibiting target blurring, severe occlusion, and small-scale objects, forming the IMC-Supplement dataset. This supplementary set comprehensively incorporates challenges like feature degradation and feature absence, rigorously testing detection models' robustness and adaptability.

The comparative evaluation encompassed a diverse set of mainstream object detection architectures, including both single-stage and two-stage detectors, as well as advanced large-backbone-based models. Performance was assessed using standard metrics such as mean Average Precision (mAP) at IoU thresholds of 0.5 and 0.75, Average Recall (AR), and inference speed (FPS). Table IV reports the results under challenging conditions—such as occlusion, small object instances, and low-contrast backgrounds—based on the IMC-Supplement dataset.

Across all detector types, a consistent degradation in accuracy is observed, with most models exhibiting a drop of over 10 points in mAP compared to their performance under standard settings. Notably, despite this degradation, **ALGC-WFEMNet(Ours)** achieves the highest mAP of **48.27**, outperforming all other baseline and large backbone models. While models like NAS-FPN and InternImage-H with SAM exhibit strong performance under normal conditions, their effectiveness diminishes more significantly in complex scenes, highlighting the superior robustness and generalization ability of the proposed method. Furthermore, **ALGC-WFEMNet** maintains a competitive inference speed (24.7 FPS) and moderate computational complexity, striking a favorable balance between accuracy and efficiency under adverse conditions.

M. Practical application testing

To validate the generalization ability of the ALGC-WFEMNet model under complex meteorological conditions, this study conducted a two-month field experiment of road target detection on typical sections such as Zhongyi Road, Furong Middle Road, and Sanyi Avenue in Changsha, China.

The ALGC-WFEMNeT system constructed in this research adopts a multimodal technical architecture, integrating the Cascade R-CNN vision model, MobileSAM segmentation algorithm, and DJI Mini 4 Pro UAV platform. The system mainly includes: Image Acquisition Module: Equipped with a 20-megapixel CMOS sensor (resolution 4000×2250), the UAV supports dynamic focus adjustment from 0-90 cm. It performs continuous image capture in monitored areas through preset aerial photography parameters. Image Loading Module: Realizes real-time transmission and preprocessing of collected data based on IoT protocols. Server Processing Module: After completing target detection operations, detection results are visualized through a host computer. Fig. 11 shows the schematic diagram of the system workflow.

In comparative experiments, we selected 50 groups of real-scenario UAV data for testing under four representative adverse weather conditions (denoted as Class 1–4): heavy rainfall, snowfall, dense fog, and low-light nighttime environments. As shown in Fig. 12, ALGC-WFEMNet consistently outperforms the baseline Cascade R-CNN across all challenging scenarios. Specifically, in Class 1 (rainfall), ALGC-WFEMNet detected 36 valid targets, outperforming the baseline's 30 and achieving a 20% improvement in detection count. In Class 2 (snowfall), it achieved 30 correct detections compared to only 20 from the baseline, marking a 50% increase in recall under obscured snowy conditions. In Class 3 (dense fog), our model successfully identified 39 targets versus 22 from the baseline, indicating a 77.3% gain and a significantly extended perceptual range under severe visibility degradation. In Class 4 (low-light nighttime), ALGC-WFEMNet achieved 36 detections, improving over the baseline's 28, and matching its own performance in Class 1, thus demonstrating consistent robustness against illumination variance and noise-induced false positives.

N. Deployment Efficiency

To assess the balance between detection accuracy and inference latency, we compared ALGC-WFEMNet against various mainstream object detection frameworks. As shown in Table V, experiments were conducted using the IMC dataset with input images resized to 800×800.

Our proposed ALGC-WFEMNet, built on the ResNet-101 backbone, achieves the highest mAP50 of **79.31%**, outperforming all other models. Its inference time of **23.31s** serves as the baseline for evaluating runtime efficiency. Compared to one-stage detectors such as NAS-FPN and M2Det, which run slightly faster (21.85s and 22.76s), ALGC-WFEMNet delivers superior detection accuracy. Two-stage counterparts like Libra R-CNN and MegDet exhibit slightly longer inference times (25.90s and 26.42s), but significantly lower accuracy. More notably, large backbone-based detectors such as Cascade R-CNN with Intern-H and InternImage-H+SAM experience a dramatic increase in latency, reaching **93.24s** and **116.55s** respectively—**4 to 5 times slower than ALGC-WFEMNet**—despite offering comparable mAP50 values. This highlights the computational inefficiency of such large-scale models in practical deployment.

To further validate the deployability of our approach on edge devices, we implemented ALGC-WFEMNet on a Raspberry Pi 4 platform.Despite limited computational resources, ALGC-WFEMNet maintains competitive performance with tolerable inference latency, significantly outperforming heavier models in terms of speed. These results confirm its suitability for real-world, resource-constrained remote sensing scenarios, such

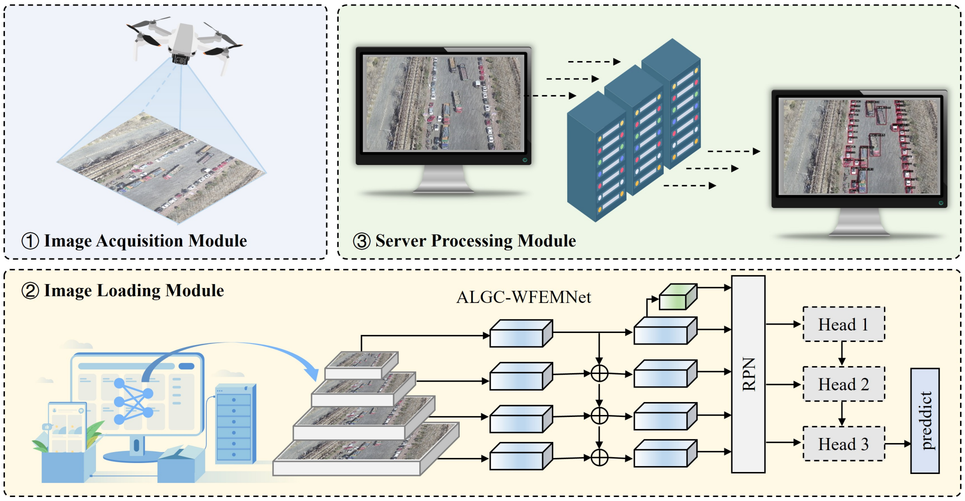


Fig. 11. Schematic diagram of the detection system of the IoT based on ALGC-WFEMNet.

TABLE IV
PERFORMANCE COMPARISON OF OBJECT DETECTION METHODS ON CHALLENGING SAMPLES.

Method	Backbone	mAP	mAP50	mAP75	AR	FPS	Parameter	GFLOPs
One-stage detectors								
SSD512 [40]	VGG16	35.21	47.96	39.82	42.97	90.7	180.73	129.13
RetinaNet [41]	ResNeXt101	45.06	56.55	50.91	49.26	68.1	55.61	12.1
RefineDet [42]	ResNet101	46.32	58.85	54.03	51.49	63.7	54.53	88.4
CornerNet [43]	Hourglass104	45.15	60.51	53.19	51.97	23.0	85.3	135
M2Det [44]	VGG16	47.51	62.26	54.39	53.95	25.1	98.9	44.14
FSAF [45]	ResNeXt101	46.59	61.48	53.83	55.31	24.9	45	15.9
NAS-FPN [46]	AmoebaNet	47.88	65.32	58.82	58.17	21.8	166.5	281.3
YOLOv3 + ASFF [47], [48]	Darknet53	39.79	52.13	45.21	54.81	30.5	55	140.69
YOLOv4 [49]	CSPDarknet53	40.31	54.12	47.09	45.78	36.0	64.36	60.52
PP-YOLO [50]	ResNet50-vd	43.32	55.32	48.63	49.74	159.2	44.93	44.71
Two-stage detectors								
Faster R-CNN [51]	VGG16	28.83	40.19	32.55	38.77	40.1	258.4	84.19
R-FCN [52]	ResNet101	36.98	47.63	41.72	51.83	22.4	59.2	46.3
FPN [53]	ResNet101	40.66	52.10	45.31	48.57	35.6	45.67	29.99
Mask R-CNN [54]	ResNet101	42.01	51.88	46.92	48.33	44.8	56.55	195.54
Libra R-CNN [55]	RseNext101	46.18	51.56	50.09	51.27	11.6	176.64	17.85
SNIP [56]	DPN-98	45.10	58.77	55.51	51.89	8.6	38.6	12.88
SINPER [57]	ResNet101	45.67	58.41	54.41	50.90	14.5	45.7	63.4
MegDet [58]	ResNet50	46.02	63.76	56.64	57.52	16.8	34.6	78.3
Cascade R-CNN [29]	ResNet101	43.87	61.59	52.38	58.48	25.6	88.16	149.48
Cascade R-CNN+MobileSAM [29], [59]	ResNet101	43.87	61.59	52.38	58.48	25.6	88.16	149.48
ALGC-WFEMNet(Ours)	ResNet101	**48.27**	**67.10**	**57.34**	**59.65**	**24.7**	**88.91**	**150.12**
Large Backbone-based detectors								
ViT-Adapter-L [60]	ViT-L	46.82	65.31	57.46	58.93	18.2	303.1	349.5
Cascade R-CNN [29]	Swin-L	47.45	63.78	58.12	59.04	17.4	246.8	298.3
Mask R-CNN [54]	ConvNeXt-XL	46.12	63.89	56.04	56.67	20.7	279.6	321.4
Cascade R-CNN [29]	Intern-H	47.38	64.71	57.22	57.53	15.8	346.2	386.1
Cascade R-CNN	InternImage-H + SAM	47.60	64.87	57.43	57.84	15.2	347.9	388.7
HRFormer-Det [61]	HRFormer-B	46.14	63.41	55.57	55.85	18.5	241.6	270.1
P2BNet-RS [62]	Swin-B	45.89	62.75	54.42	54.83	17.6	222.5	242.9
RFLA-FPN++ [63]	ResNeSt200	46.08	62.90	55.12	55.28	20.2	208.3	237.1
RSFormer [64]	Mamba-B	46.71	64.72	56.66	56.92	18.1	268.7	312.8

as UAV-based environmental monitoring and rapid disaster response.

IV. CONCLUSIONS

In this paper, prior to conducting formal experiments, we constructed a dataset of road scenes captured by drones under adverse weather conditions. Based on this dataset, we

TABLE V
INFERENCE TIME AND ACCURACY COMPARISON BETWEEN ALGC-WFEMNET AND REPRESENTATIVE DETECTION FRAMEWORKS. TIME RATIO DENOTES THE RELATIVE INFERENCE TIME COMPARED TO ALGC-WFEMNET. OUR METHOD ACHIEVES THE HIGHEST MAP50 WHILE MAINTAINING ACCEPTABLE LATENCY.

Framework	Backbone	mAP	mAP50	Time (s)	Time Ratio
One-stage detectors					
NAS-FPN	AmoebaNet	60.35	78.03	21.85	0.94×
M2Det	VGG16	59.24	73.52	22.76	0.98×
Two-stage detectors					
Libra R-CNN	ResNet101	58.24	64.13	25.90	1.11×
MegDet	ResNet50	58.31	77.14	26.42	1.13×
ALGC-WFEMNet (Ours)	ResNet101	**60.48**	**79.31**	**23.31**	**1.00**×
Large Backbone-based detectors					
Cascade R-CNN	InternImage-H+SAM	60.02	78.56	116.55s	5.00×
Cascade R-CNN	Intern-H	59.94	78.44	93.24s	4.00 ×

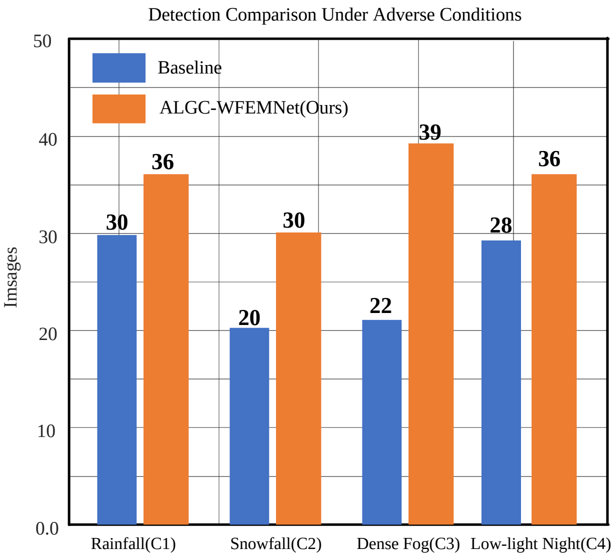


Fig. 12. Quantitative comparison of detection performance under four representative adverse weather conditions on the IMC dataset. Each class includes 50 UAV-captured images representing real-world scenarios. Compared to the baseline Cascade R-CNN, our proposed ALGC-WFEMNet consistently outperforms across all settings. In Class 1 (rainfall), the model achieves a 23% improvement in detection box localization. Under Class 2 (snowfall), it recovers 35% of the previously missed targets. In Class 3 (dense fog), ALGC-WFEMNet extends the effective detection capability to 1.8× that of the baseline. In Class 4 (low-light nighttime), it reduces false detections, maintaining the same number of true detections (36 vs. 28) while improving precision. These results highlight the model's robustness in complex environmental conditions and its strong generalization ability for real-time UAV-based remote sensing applications.

investigated the challenges associated with road inspection tasks using drones in complex weather environments, within the framework of remote sensing object detection. Our analysis highlighted the limitations of traditional detection models, particularly in effectively integrating multi-scale information. To address these limitations, we proposed ALGC, a novel remote sensing object detection model that enhances multi-scale feature fusion through the use of 1×1 convolutions within the FPN structure. Furthermore, we introduced the WFEM module, which amplifies the contrast between pixel-level feature values and enhances the network's ability to restore discriminative feature representations.

Experimental results across various detectors, backbone networks, and datasets validate both the effectiveness and generalization capability of ALGC. However, we observed a significant performance degradation when using Vision Transformers (ViTs) as backbones. This is likely attributed to their reliance on large-scale datasets to perform global feature learning, which makes them less effective in processing UAV imagery characterized by local variations and substantial noise in adverse weather conditions. In such scenarios, the lack of spatial inductive bias in ViTs hinders their ability to focus on relevant local structures.

Taken together, our work introduces an ALGC-enhanced multi-scale fusion framework and a WFEM-driven weather-aware representation module, supported by a dedicated adverse-weather UAV dataset, forming a complete solution for robust remote sensing object detection in challenging environments. To address these challenges ulteriorly, future work will explore hybrid ViT-CNN architectures that incorporate spatial biases and facilitate more effective multi-scale representation learning, thereby improving robustness in complex remote sensing environments. In addition, the recent emergence of lightweight foundation models such as MobileSAM presents new opportunities. MobileSAM offers fast inference capabilities and strong generalization across tasks, making it well-suited for deployment on resource-constrained aerial platforms. Its ability to rapidly produce accurate segmentation under limited computational budgets may complement existing detection pipelines by providing efficient region proposals or enhancing feature localization. This is particularly valuable in adverse weather conditions, where rapid response and real-time adaptability are critical. Incorporating such efficient foundation models into the remote sensing detection framework could substantially improve the adaptability and responsiveness of UAV-based inspection systems in dynamic and challenging environments.

REFERENCES

[1] J. Zhou, Y. Liu, B. Peng, L. Liu, and X. Li, "Madinet: Mamba diffusion network for sar target detection," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2025.

[2] j. Zhou, Y. Liu, L. Liu, W. Li, B. Peng, Y. Song, G. Kuang, and X. Li, "Fifty years of sar automatic target recognition: The road forward," *arXiv preprint arXiv:2501.22159*, 2025. [Online]. Available: https://arxiv.org/abs/2509.22159

[3] J. Li, K. Zheng, Z. Li, L. Gao, and X. Jia, "X-shaped interactive autoencoders with cross-modality mutual learning for unsupervised hyperspectral image super-resolution," *IEEE transactions on geoscience and remote sensing*, vol. 61, pp. 1–17, 2023.

[4] X. Zhong, J. Zhan, Y. Xie, L. Zhang, G. Zhou, M. Liang, K. Yang, Z. Guo, and L. Li, "Adaptive deformation-learning and multiscale-integrated network for remote sensing object detection," *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[5] J. Chen, L. Liu, W. Deng, Z. Liu, Y. Liu, Y. Wei, and Y. Liu, "Refining pseudo labeling via multi-granularity confidence alignment for unsupervised cross domain object detection," *IEEE Transactions on Image Processing*, 2025.

[6] Z. Zheng, Y. Zhong, A. Ma, X. Han, J. Zhao, Y. Liu, and L. Zhang, "Hynet: Hyper-scale object detection network framework for multiple spatial resolution remote sensing imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 166, pp. 1–14, 2020.

[7] J. Zhou, C. Xiao, B. Peng, Z. Liu, L. Liu, Y. Liu, and X. Li, "Diffdet4sar: Diffusion-based aircraft target detection network for sar images," *IEEE Geoscience and Remote Sensing Letters*, 2024.

[8] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," pp. 21–37, 2016.

[9] X. Liang, J. Zhang, L. Zhuo, Y. Li, and Q. Tian, "Small object detection in unmanned aerial vehicle images using feature fusion and scaling-based single shot detector with spatial context analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 6, pp. 1758–1770, 2020.

[10] G. Wang, Y. Zhuang, H. Chen, X. Liu, T. Zhang, L. Li, S. Dong, and Q. Sang, "Fsod-net: Full-scale object detection from optical remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–18, 2022.

[11] Q. Li, L. Mou, Q. Liu, Y. Wang, and X. X. Zhu, "Hsf-net: Multiscale deep feature embedding for ship detection in optical remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 12, pp. 7147–7161, 2018.

[12] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[13] L. Hou, K. Lu, and J. Xue, "Refined one-stage oriented object detection method for remote sensing images," *IEEE Trans. Image Process.*, vol. 31, pp. 1545–1558, 2022.

[14] W. Zhang, L. Jiao, Y. Li, Z. Huang, and H. Wang, "Laplacian feature pyramid network for object detection in vhr optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.

[15] J. Fu, X. Sun, Z. Wang, and K. Fu, "An anchor-free method based on feature balancing and refinement network for multiscale ship detection in sar images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 2, pp. 1331–1344, 2021.

[16] J. Liu, J. Zhan, J. Zhang, J. Chen, Y. Song, L. Tang, L. Zhou, C. Du, Y. Wei, and Y. Guo, "Robust scale fusion and edge-aware feature attention network for remote sensing uav road detection under harsh weather," *Results in Engineering*, p. 106172, 2025.

[17] R. Wang, H. Zhao, Z. Xu, Y. Ding, G. Li, Y. Zhang, and H. Li, "Real-time vehicle target detection in inclement weather conditions based on yolov4," *Frontiers in Neurorobotics*, vol. 17, 2023.

[18] L. Liu, S. Sun, S. Zhi, F. Shi, Z. Liu, J. Heikkila, and Y. Liu, "A causal adjustment module for debiasing scene graph generation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–16, 2025.

[19] 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*, 2020.

[20] Z. Zhu, X. Li, J. Zhai, and H. Hu, "Podb: A learning-based polarimetric object detection benchmark for road scenes in adverse weather conditions," *Information Fusion*, vol. 108, 2024.

[21] J. Yao, X. Fan, B. Li, and W. Qin, "Adverse weather target detection algorithm based on adaptive color levels and improved yolov5," *Sensors*, vol. 22, no. 21, p. 8577, 2022.

[22] M. Jeon, J. Seo, and J. Min, "Da-raw: Domain adaptive object detection for real-world adverse weather conditions," *arXiv*, 2023. [Online]. Available: https://arxiv.org/abs/2309.08152

[23] W. Liu, G. Ren, R. Yu, S. Guo, J. Zhu, and L. Zhang, "Image-adaptive yolo for object detection in adverse weather conditions," *arXiv*, 2021. [Online]. Available: https://arxiv.org/abs/2112.08088

[24] T. Sharma, B. Debaque, N. Duclos, A. Chehri, B. Kinder, and P. Fortier, "Deep learning-based object detection and scene perception under bad weather conditions," *Electronics*, vol. 11, no. 4, p. 563, 2022.

[25] M. Hildebrand, A. Brown, S. Brown, and S. L. Waslander, "Assessing distribution shift in probabilistic object detection under adverse weather," *IEEE Access*, 2023.

[26] J. Zhan, Y. Xie, J. Guo, Y. Hu, G. Zhou, W. Cai, Y. Wang, A. Chen, L. Xie, M. Li *et al.*, "Dgpf-renet: A low data dependence network with low training iterations for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–21, 2023.

[27] J. Li, K. Zheng, L. Gao, Z. Han, Z. Li, and J. Chanussot, "Enhanced deep image prior for unsupervised hyperspectral image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[28] J. Li, K. Zheng, L. Gao, L. Ni, M. Huang, and J. Chanussot, "Model-informed multistage unsupervised network for hyperspectral image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–17, 2024.

[29] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6154–6162.

[30] F. Li and F. Li, "Bag of tricks for fabric defect detection based on cascade r-cnn," *Textile Research Journal*, vol. 91, no. 5-6, pp. 599–612, 2021.

[31] Y. Liu, Y. Zhao, X. Zhang, X. Wang, C. Lian, J. Li, P. Shan, C. Fu, X. Lyu, L. Li *et al.*, "Mobilesam-track: lightweight one-shot tracking and segmentation of small objects on edge devices," *Remote Sensing*, vol. 15, no. 24, p. 5665, 2023.

[32] J. Zhang, C. Li, Y. Yin, J. Zhang, and M. Grzegorzek, "Applications of artificial neural networks in microorganism image analysis: a comprehensive review from conventional multilayer perceptron to popular convolutional neural network and potential visual transformer," *Artificial Intelligence Review*, vol. 56, no. 2, pp. 1013–1070, 2023.

[33] Z. Tan, Y. Wu, Q. Liu, Q. Chu, L. Lu, J. Ye, and N. Yu, "Exploring the application of large-scale pre-trained models on adverse weather removal," *IEEE Transactions on Image Processing*, 2024.

[34] Y. Yang, A. I. Aviles-Rivero, H. Fu, Y. Liu, W. Wang, and L. Zhu, "Video adverse-weather-component suppression network via weather messenger and adversarial backpropagation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 13 200–13 210.

[35] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu *et al.*, "Mmdetection: Open mmlab detection toolbox and benchmark," *arXiv preprint arXiv:1906.07155*, 2019.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[37] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[38] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 568–578.

[39] W. Yu, M. Luo, P. Zhou, C. Si, Y. Zhou, X. Wang, J. Feng, and S. Yan, "Metaformer is actually what you need for vision," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 819–10 829.

[40] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 21–37.

[41] Y. Wang, C. Wang, H. Zhang, Y. Dong, and S. Wei, "Automatic ship detection based on retinanet using multi-resolution gaofen-3 imagery," *Remote Sensing*, vol. 11, no. 5, p. 531, 2019.

[42] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Single-shot refinement neural network for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4203–4212.

[43] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 734–750.

[44] Q. Zhao, T. Sheng, Y. Wang, Z. Tang, Y. Chen, L. Cai, and H. Ling, "M2det: A single-shot object detector based on multi-level feature pyramid network," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 9259–9266.

[45] N. Kumar Sharma, B. Kalyani Immadisetty, A. Govina, R. Chandra Reddy, and P. Choubey, "Corn leaf disease detection using resnext50, resnext101, and inception v3 deep neural networks," in *Machine Vision and Augmented Intelligence: Select Proceedings of MAI 2022*. Springer, 2023, pp. 303–313.

[46] G. Ghiasi, T.-Y. Lin, and Q. V. Le, "Nas-fpn: Learning scalable feature pyramid architecture for object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 7036–7045.

[47] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[48] M. Qiu, L. Huang, and B.-H. Tang, "Asff-yolov5: Multielement detection method for road traffic in uav images based on multiscale feature fusion," *Remote Sensing*, vol. 14, no. 14, p. 3498, 2022.

[49] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[50] X. Long, K. Deng, G. Wang, Y. Zhang, Q. Dang, Y. Gao, H. Shen, J. Ren, S. Han, E. Ding *et al.*, "Pp-yolo: An effective and efficient implementation of object detector," *arXiv preprint arXiv:2007.12099*, 2020.

[51] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[52] J. Dai, Y. Li, K. He, and J. Sun, "R-fcn: Object detection via region-based fully convolutional networks," *Advances in neural information processing systems*, vol. 29, 2016.

[53] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," 2017. [Online]. Available: https://arxiv.org/abs/1612.03144

[54] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[55] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, and D. Lin, "Libra r-cnn: Towards balanced learning for object detection," 2019. [Online]. Available: https://arxiv.org/abs/1904.02701

[56] B. Singh and L. S. Davis, "An analysis of scale invariance in object detection - snip," 2018. [Online]. Available: https://arxiv.org/abs/1711.08189

[57] B. Singh, M. Najibi, and L. S. Davis, "SNIPER: Efficient multi-scale training," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 31, 2018, pp. 9333–9343.

[58] C. Peng, T. Xiao, Z. Li, Y. Jiang, X. Zhang, K. Jia, G. Yu, and J. Sun, "Megdet: A large mini-batch object detector," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6181–6189.

[59] C. Zhang, D. Han, Y. Qiao, J. U. Kim, S.-H. Bae, S. Lee, and C. S. Hong, "Faster segment anything: Towards lightweight sam for mobile applications," 2023. [Online]. Available: https://arxiv.org/abs/2306.14289

[60] Z. Chen, Y. Duan, W. Wang, J. He, T. Lu, J. Dai, and Y. Qiao, "Vision transformer adapter for dense predictions," *arXiv preprint arXiv:2205.08534*, 2022.

[61] Y. Yuan, R. Fu, L. Huang, W. Lin, C. Zhang, X. Chen, and J. Wang, "Hrformer: High-resolution vision transformer for dense predict," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, 2021, pp. 7281–7293.

[62] P. Chen, X. Yu, X. Han, N. Hassan, K. Wang, J. Li, J. Zhao, H. Shi, Z. Han, and Q. Ye, "Point-to-box network for accurate object detection via single point supervision," in *European Conference on Computer Vision*. Springer, 2022, pp. 51–67.

[63] C. Xu, J. Wang, W. Yang, H. Yu, L. Yu, and G.-S. Xia, "Rfla: Gaussian receptive field based label assignment for tiny object detection," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*. Springer, 2022, pp. 526–543.

[64] T. Gao, Y. Wen, K. Zhang, P. Cheng, and T. Chen, "Towards an effective and efficient transformer for rain-by-snow weather removal," 2023. [Online]. Available: https://arxiv.org/abs/2304.02860

[65] A. Chandio, G. Gui, T. Kumar, I. Ullah, R. Ranjbarzadeh, A. M. Roy, A. Hussain, and Y. Shen, "Precise single-stage detector," *arXiv preprint arXiv:2210.04252*, 2022.

[66] C. Zhu, Y. He, and M. Savvides, "Feature selective anchor-free module for single-shot object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 840–849.

[67] Y. Wang, X. Zhang, T. Yang, and J. Sun, "Anchor detr: Query design for transformer-based detector," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, no. 3, 2022, pp. 2567–2575.