AccuLiteFastNet: A Remote Sensing Object Detection Model Combining High Accuracy, Lightweight Design, and Fast Inference Speed

Ruihan Bai, Feng Shen*, Mingkang Wang, Jiahui Lu, and Zhiping Zhang

***Abstract*—This letter proposes AccuLiteFastNet, a novel object detection model tailored for aerial remote sensing scenarios constructed on the YOLOv8. AccuLiteFastNet is designed to enhance the precision in detecting small-size objects while maintaining a lightweight model framework and supporting fast inference speeds for deployment. Specifically, we use the Normalized Gaussian Wasserstein Distance (NGWD) as the loss metric to evaluate the similarity between predicted and ground-truth object bounding boxes, ensuring uniform sensitivity for objects of different sizes. Moreover, we replace the model backbone's last feature extraction module with the self-designed Multi-Scale FasterNet (MSFN) module to reduce the model complexity while providing an expanded receptive field. Finally, we propose a novel module named ContA-C2f specifically designed to evaluate the informativeness of each pixel in the image. Across three open-source remote sensing datasets, AccuLiteFastNet demonstrates a significant improvement over the original YOLOv8, achieving a 4.5% increase in mean Average Precision (mAP). Comparisons with existing object detection models, AccuLiteFastNet strikes an optimal balance with high inference speed (58.14 FPS) and good accuracy (87.5% mAP) at lower computational costs (8.3 GFLOPs).**

***Index Terms*—Object detection, YOLOv8, Remote sensing.**

I. Introduction

REMOTE sensing technology utilizes devices like unmanned aerial vehicles (UAVs) and satellites to collect image data from the earth's surface. Due to its ability to provide extensive and continuous coverage of terrestrial information, it is widely used in climate change research, military reconnaissance, urban planning, and disaster monitoring. However, compared to ordinary natural scene images, remote sensing image has its unique characteristics. Specifically, the objects within an image occupy a minimal pixel area, with most of the space filled by a complicated background. The scarcity of pixels limits the visual information accessible, which in turn hinders the object detection models' ability to extract distinctive features. Furthermore, the features of a target can easily become obscured by background noise and other visual elements, such as colors, textures, or shapes that resemble the target. Consequently, the objects become susceptible to interference from the complex background information. With the development of deep learning, many researchers have started to use target detection models such as Convolutional Neural Networks (CNN) to learn the features of the targets within remote sensing images. Chen et al. [1] proposed a deep learning-based approach for identifying objects in remote sensing. Their study overviews several critical aspects of microscopic item identification: multi-scale feature, contextual information, and super-resolution. Liu et al. [2] introduced a method involving multi-block SSDs that includes sub-layers for the detection and enhancement of local context information. Comparative analysis of test results between multi-block SSDs and traditional SSDs revealed that the suggested model significantly improved the detection rate of small objects by 23.2%. Bosquet et al. [3] introduced STDNet and ConvNet as approaches for identifying tiny objects smaller than 16×16 pixels based on regional concepts. Zheng et al. [4] introduced a novel HyNet framework for large-scale target recognition in MSR remote sensing imaging. Zhang et al. [5] introduced SuperYOLO, a method offering rapid and precise object detection in remote sensing imagery. It adeptly manages small, multi-scale objects within intricate backgrounds by employing an innovative combination of multimodal data fusion and super-resolution strategies. Zhang et al. [6] introduced a wavelet multiscale attention mechanism that emphasizes objects against backgrounds, improving detection capabilities across different neural network architectures. Liu et al. [7] introduced the Foreground Refinement Network (FoRDet), designed to improve object detection in remote sensing by concentrating on the details of the foreground and simplifying background complexity. DSDL-Net addresses the challenge of accurately detecting multiscale objects against complex backgrounds in remote sensing by integrating a multi-receptive field fusion module for semantic learning and an adaptive fusion network for detail preservation [8].

While certain research efforts concentrate on enhancing the accuracy of detection models, they may unintentionally neglect the importance of reducing the models' computational load. Additionally, the models proposed in certain studies might exhibit a reduction in the inference speed, which may be attributed to the transformer-based contextual modules introduced by related studies. Such modules tend to perform a grid division of the input features and look for correlations between them, thus potentially increasing the computational complexity. Remotely sensed data is often used in real-time or near real-time decision-making applications such as surface change monitoring and disaster response, so the real-time characteristic of the models is critical. In this research, we proposed AccuLiteFastNet, a novel lightweight yet relatively high-precision target detection network tailored for aerial remote sensing scenarios. AccuLiteFastNet is built upon the YOLOv8 network framework, the specific design details of the model include the following three aspects:

• Use the Normalized Gaussian Wasserstein Distance (NGWD) loss function instead of the CIoU loss function to evaluate the similarity between predicted and ground-truth object bounding boxes, ensuring consistent sensitivity across objects of varying sizes.

• Due to high similarity across different channels, replace the model backbone's last feature extraction module with the self-designed Multi-Scale FasterNet (MSFN) module to reduce the model complexity while providing an expanded receptive field.

• Propose a novel module named ContA-C2f that augments the C2f module with the Context Aggregation block, specifically designed to evaluate the informativeness of each pixel in the image.

The proposed model AccuLiteFastNet exhibits higher precision on three open-source datasets than other object detection models. In addition, the model is designed to be lightweight and provides fast inference speed.

II. Method

A. Overall of the YOLOv8 model

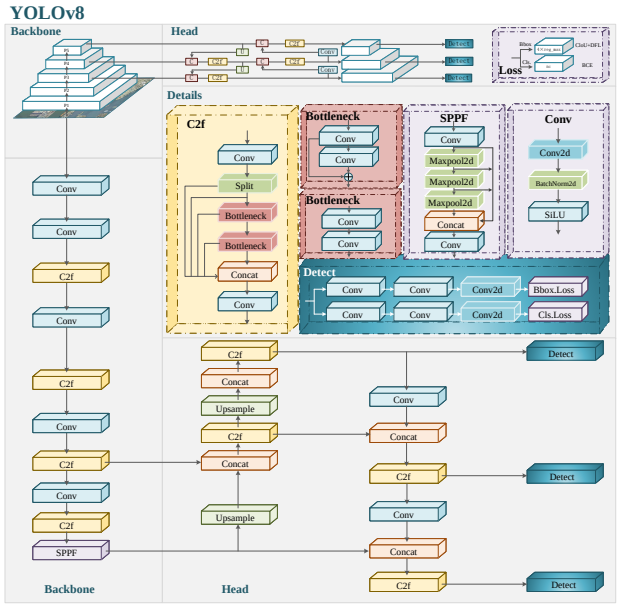

Fig. 1. The overall of the YOLOv8.

YOLOv8 (Fig. 1) is a State-of-the-art (SOTA) object detection model that Ultralytics released on January 10, 2023. The backbone of YOLOv8 primarily comprises the C2f module. The architecture of the C2f module integrates two parallel gradient flow branches, facilitating a more robust gradient information flow.

B. Normalized Gaussian Wasserstein Distance(NGWD)

The loss function of YOLOv8 for object detection bounding boxes has two parts. The first portion is Distribution Focal Loss (DFL) [9], which uses cross-entropy as an optimization mechanism to concentrate the network's predictive distribution closer to the label values. Another portion calculates the intersection over Union (IoU) between the predicted and true bounding boxes, utilizing the CIoU loss metric. The CIoU loss function displays varied sensitivity towards objects of different sizes. Small-sized objects, having fewer image pixels than regular-sized ones, demonstrate substantial fluctuations in their CIoU value with minor positional alterations of the predicted bounding boxes. However, the same positional changes for regular-sized objects produce minimal variations in the CIoU value. The sensitivity of CIoU value on small-size objects causes the labels of small objects to become opposite easily during the positive and negative sample assignment, leading to positive and negative samples with similar features. This condition complicates the convergence of the object detection network during training.

In this letter, the CIoU loss function used by YOLOv8 is replaced with the Normalized Gaussian Wasserstein Distance [10]. The idea is to convert the bounding box of the object into a two-dimensional Gaussian distribution and then measure the similarity between the two distributions using the Wasserstein distance. Specifically, for a bounding box R (c_x, c_y, w, h), c_x, c_y, w, h denote the bounding box's center coordinates, width, and height, respectively. The equation for its inner tangent ellipse is:

$$\frac{(x-c_x)^2}{(\frac{w}{2})^2}+\frac{(y-c_y)^2}{(\frac{h}{2})^2}=1 \tag{1}$$

where (c_x, c_y) represents the center coordinate of the ellipse, and (w/2, h/2) represents the length along the x-axis and y-axis.

The probability density function of a two-dimensional Gaussian distribution could be expressed as:

$$f(X|\mu,\Sigma)=\frac{exp(-\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu))}{2\pi\,|\Sigma|^{\frac{1}{2}}} \tag{2}$$

where X is the position variable, and (μ, Σ) is the mean vector and the covariance matrix.

Since equation (1) can be converted into the form of equation (3), the inner tangent ellipse is a distribution contour of the 2D Gaussian distribution N(μ, Σ), μ and Σ is shown in equation (4).

$$(X-\mu)^T\Sigma^{-1}(X-\mu)=1 \tag{3}$$

$$E(c_x,c_y,w/2,h/2)\sim N(\mu,\Sigma)|\mu=\begin{bmatrix}c_x\\c_y\end{bmatrix},\Sigma=\begin{bmatrix}\frac{w^2}{4}&0\\0&\frac{h^2}{4}\end{bmatrix} \tag{4}$$

For two Gaussian distributions $N_1(\mu_1, \Sigma_1)$ and $\mu_2(m_2, \Sigma_2)$, the Wasserstein distance between N_1 and N_2 can be defined as:

$$W_2^2(N_1,N_2)=\left\|([c_{x1},c_{y1},\frac{w_1}{2},\frac{h_1}{2}]^T,[c_{x1},c_{y1},\frac{w_1}{2},\frac{h_1}{2}]^T)\right\|_F^2 \tag{5}$$

where $||.||_F$ is Frobenius norm.

The research suggests that the Normalized Gaussian Wasserstein Distance loss function maintains consistent sensitivity

across objects of different sizes, exhibiting smooth value changes for smaller objects. This method can provide a more detailed depiction of the object's spatial distribution, thus improving object detection and localization effectiveness in numerous computer vision.

C. Multi-Scale FasterNet (MSFN)

In many research studies, it has been observed that feature maps exhibit a high degree of commonality or similarity across various channels. As shown in Fig. 2, the feature maps extracted from the backbone part of the YOLOv8 share similarities among different channels (red, purple, and green boxes).

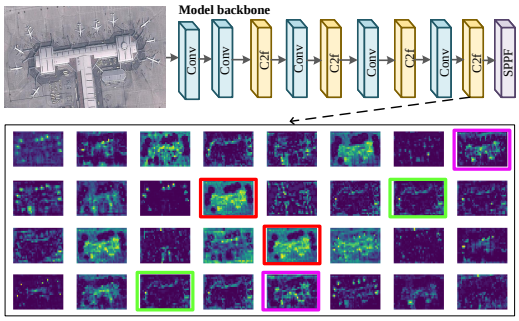

Fig. 2. Visualization of feature maps in an intermediate layer of the YOLOv8.

Partial convolution (PConv) [11] is an innovative convolution operation designed to extract spatial features efficiently while minimizing redundancy in computations and memory access. Unlike other convolution operations (traditional convolution utilizes all input feature map channels; Depthwise Convolution partitions the input feature map into multiple groups, each subject to an independent convolution operation), PConv selectively convolves a subset of channels, leaving the rest intact. The PConv operation lessens the redundancy of channel feature information during computation, thereby boosting computational efficiency. Furthermore, PConv can retain pertinent information, enhancing feature representation capabilities without compromising computational efficiency. To optimally exploit the information available in all channels, the FasterNet block supplements the PConv with a pointwise convolution (PW-Conv). The effective receptive field on the input feature map visually resembles a T-shaped convolution, which, unlike the standard convolution that evenly processes convolution patches, places more emphasis on the central location. The related research affirms the value of the T-shaped receptive field by quantifying the importance of each position via calculation of the Frobenius norm. The FasterNet block (as depicted in Fig. 3(a)) with the T-shaped receptive field principle focuses on the central position and facilitates a more comprehensive understanding of contextual information within the input image.

This study proposes a Multi-Scale FasterNet (MSFN) module (as depicted in Fig. 3(a), whose core component is the Multi-Scale FastNet block(as depicted in Fig. 3(b). The Multi-Scale FastNet block builds upon the original FasterNet block (as depicted in Fig. 3(c)). It incorporates an array of PConv with varying kernel sizes, generating accordingly T-shaped convolutions with varying sizes that enlarge the receptive field. This design enhances the model's spatial feature extraction capabilities by allowing it to gather information at multiple scales, leading to a richer understanding of the input feature. Crucially, the PConv mechanism selectively convolves specific channels, which minimizes redundancy and ensures computational resources are used more efficiently. This characteristic makes the MSFN module enlarge the receptive field without significantly increasing the model's parameters, which is essential for deployment on edge devices where computational efficiency is paramount. In summary, the MSFN module represents a balanced approach to enhancing neural network capability, providing an expanded receptive field and detailed feature representation while maintaining an efficient architecture. The MSFN module helps identify which positions are more critical for feature extraction, further optimizing the convolutional operations. For our proposed AccuLiteFastNet, we replace the last C2f module in the backbone part with the MSFN module.

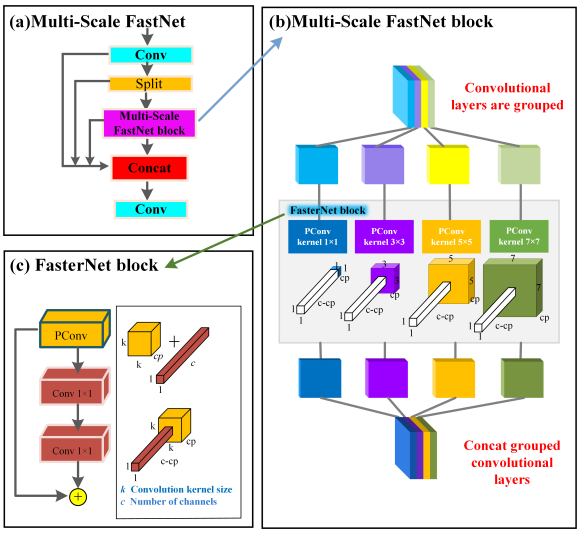

Fig. 3. The overall of the proposed MSFN module.

D. Context Aggregation-C2f (ContA-C2f)

In the analysis of remote-sensing images, we observe that objects often occupy only a tiny portion of the image area, leaving large areas of background information that may not be relevant. The conventional design does not discern between object and background information and may inadvertently incorporate excessive uninformative background features. To address this challenge, this letter proposes a novel module named ContA-C2f that augments the C2f module with the Context Aggregation block [12], specifically designed to evaluate the informativeness of each pixel in the image. For the structure of the Context Aggregation block (as shown in Fig. 4(a)), two dedicated branches generate attention and feature maps, while a separate third branch captures context maps. The attention and feature maps are merged through a matrix multiplication process that adaptively adjusts the input

features, thereby maintaining crucial global information. The product of this matrix multiplication is subsequently combined with the context maps through an element-wise addition to yield the enhanced context refinement maps.

Overall, the Context Aggregation block principle is straightforward: if a pixel's features are deemed sufficiently informative, there is a minimal necessity to aggregate features from other spatial positions. The block maintains an adaptive balance between incorporating essential global context and preserving distinct local features, enhancing the model's ability to discern features within a larger context while retaining detailed local differences. As shown in Fig. 4(b), the ContA-C2f module proposed in this study is a Context Aggregation module added after the C2f module to improve the ability of the model to learn contextual information.

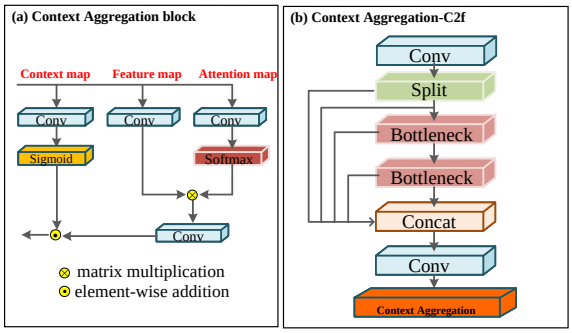

Fig. 4. The overall of the proposed ContA-C2f module.

III. Experiment Data Introduction

This letter utilized a combined experimental dataset comprising several open-source collections, including the NWPU VHR-10 remote sensing images, RSOD remote sensing images, and VisDrone UAV images. As the primary component of the experimental dataset, the NWPU VHR-10 dataset contains ten categories: aircraft, ships, storage tanks, baseball fields, tennis courts, basketball courts, ground tracks, ports, bridges, and vehicles. Despite this diversity, the NWPU VHR-10 dataset's amount is somewhat limited, with approximately 600 samples. The aircraft and storage tank classes from the RSOD dataset and the vehicle class from the VisDrone dataset were incorporated into the NWPU VHR-10 dataset to enrich the experimental datasets. Therefore, the final dataset was expanded to approximately 1500 samples.

IV. Experimental Results and Analysis

A. Evaluation metric and Experimental hyperparameter

Considering the variability in image sizes, this letter adopted a relative value-based classification where the size of a target is determined by the square root of the ratio of the target's pixel area to the total pixel area of the image. Based on this criterion, targets are classified as follows: Small targets: occupying less than 5% of the total image area, Medium targets: occupying between 5% and 10% of the total image area, Large targets: occupying more than 10% of the total image area. PR curves show how precision and recall change with confidence thresholds. The area under these curves is the AP, and the average AP across categories is the mAP. Thus, the mAP for small, medium, and large targets corresponds to mAP_s, mAP_m, and mAP_l, respectively. The hyperparameters for the experiment are summarized in Table 1.

TABLE I
Experimental Hyperparamete

Experimental hyperparameter	Value
lr0	0.01
lrf	0.01
Optimizer	Adam
Batch size	55
Training epoch	100
Image input size	640

B. Ablation experimention

TABLE II
The Results of the Ablation Experiments

Model	mAP_s	mAP_m	mAP_l	mAP
1 baseline (YOLOv8-n)	0.425	0.416	0.497	0.830
2 baseline+NGWD	0.436	0.417	0.512	0.863
3 baseline+NGWD+MSFN	0.437	0.390	0.525	0.871
4 baseline+NGWD+MSFN+ContA-C2f	0.442	0.425	0.537	0.875

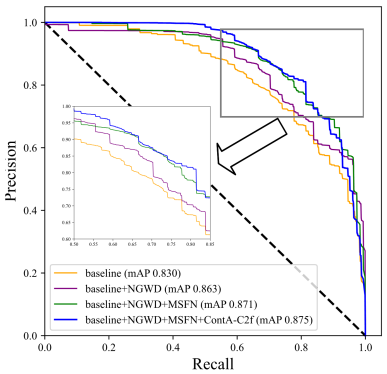

Fig. 5. The comparison of PR curves with different models.

In this section, the performance of the proposed improvement is assessed systematically via ablation studies on experimental datasets. Table 2 summarizes Yolov8 models with different improvement strategies: NGWD, MSFN, and ContA-C2f. Based on the results from the ablation studies, the three modifications implemented have led to enhancements in the model's performance. Specifically, there was an improvement of 1.7% in mAP_s, 0.9% in mAP_m, and 4% in mAP_l. Furthermore, the PR curves illustrated in Fig. 5 indicate that the modified model leans more towards the upper right corner, signifying a larger Area Under the Curve (AUC) value.

To effectively illustrate the improved algorithm's superiority, this letter performed testing on several datasets of images from the three open-source dataset. As depicted in Fig. 6, and in conjunction with the mAP scores for large, medium, and small targets, AccuLiteFastNet has substantially reduced the number of missed detections compared to the original model across three datasets.

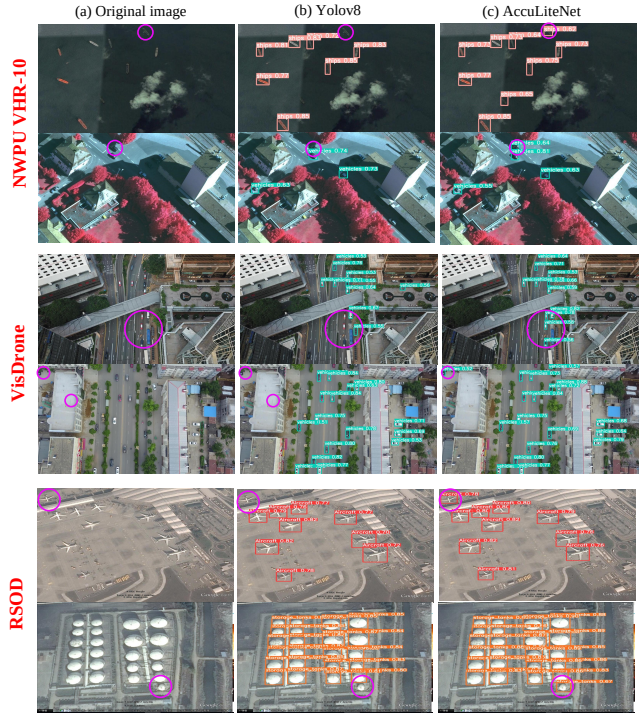

Fig. 6. Detection results of YOLOv8-n and AccuLiteFastNet across three open-source remote sensing datasets.

C. Compare with other models

TABLE III
COMPARATIVE PERFORMANCE RESULTS OF DIFFERENT MODELS

Model	mAP$_s$	mAP$_m$	mAP$_l$	mAP	GFLOPs	FPS
YOLOv5-s	0.416	0.345	0.589	0.830	15.8	57.14
YOLOv5-n	0.343	0.334	0.448	0.702	4.2	60.98
YOLOX-s	0.474	0.354	0.428	0.829	21.6	51.81
YOLOX-n	0.409	0.295	0.431	0.769	5.6	53.76
YOLOv7-tiny	0.501	0.304	0.389	0.825	13.1	52.36
YOLOv8-n	0.425	0.416	0.497	0.830	8.1	56.18
SuperYOLO	0.634	0.269	0.338	0.812	10.03	58.42
Faster RCNN	0.435	0.383	0.663	0.907	206.71	26.34
Cascade RCNN	0.452	0.421	0.651	0.913	234.49	23.41
AccuLiteFastNet	0.442	0.425	0.537	0.875	8.3	58.14

In this section, the letter compared the proposed AccuLiteFastNet with other models like YOLOv5-n, YOLOv5-s, YOLOX-n, YOLOX-s, YOLOv7-tiny, SuperYOLO [5], Faster RCNN [10], and Cascade RCNN [10]. All experimental results are shown in Table 3. While Cascade RCNN offers the highest accuracy (0.913 mAP), its heavy computational load limits its speed. Despite its leading speed (60.98 FPS), YOLOv5-n sacrifices overall accuracy. In comparison, AccuLiteFastNet strikes an optimal balance with high inference speed (58.14 FPS) and good accuracy (0.875 mAP) at lower computational costs (8.3 GFLOPs). The Fig. 7 presents a comprehensive analysis of various object detection models by plotting their performance in terms of FPS on the x-axis against the mAP on the y-axis. Each scatter point's size reflects the model's parameter count. The result demonstrate that the AccuLiteFastNet model demonstrates an optimal balance.

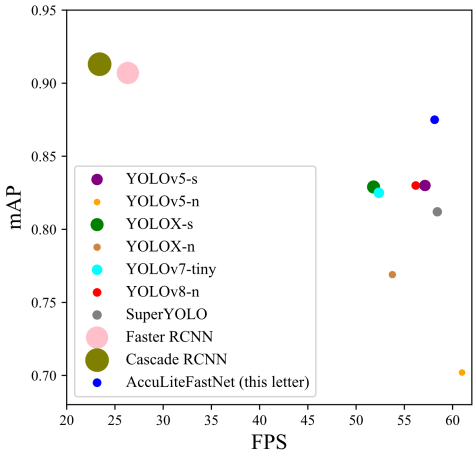

Fig. 7. Comparative analysis of model inference speed, accuracy, and complexity.

V. CONCLUSION

This letter proposed AccuLiteFastNet, a novel object detection network optimized for aerial remote sensing constructed on the YOLOv8 network framework. The result indicates that the AccuLiteFastNet model displayed superior detection precision, model complexity, and inference speed performance.

REFERENCES

[1] G. Chen, H. Wang, K. Chen, Z. Li, Z. Song, Y. Liu, W. Chen, and A. Knoll, "A Survey of the Four Pillars for Small Object Detection: Multi-scale Representation, Contextual Information, Super-Resolution, and Region Proposal," *IEEE T. Syst Man Cy-S.*, vol. 52, no. 2, pp. 936–953, 2022.

[2] S. Liu, D. Huang, and Y. Wang, "Receptive Field Block Net for Accurate and Fast Object Detection," 2018, *arXiv:1711.07767*.

[3] B. Bosquet, M. Mucientes, and V. M. Brea, "STDnet: Exploiting High Resolution Feature Maps for Small Object Detection," *Eng. Appl Artif Intel.*, vol. 91, pp. 103615, 2020.

[4] 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. Photogramm.*, vol. 166, pp. 1–14, 2020.

[5] J. Zhang, J. Lei, W. Xie, Z. Fang, Y. Li, and Q. Du, "SuperYOLO: Super resolution assisted object detection in multimodal remote sensing imagery," *IEEE T GEOSCI REMOTE.*, vol. 61, pp. 1–15, 2023.

[6] J. Bai, J. Ren, Y. Yang, Z. Xiao, W. Yu, and V. Havyarimana, "Object detection in large-scale remote-sensing images based on time-frequency analysis and feature optimization," *IEEE T GEOSCI REMOTE.*, vol. 60, pp. 1–16, 2021.

[7] T. Zhang, X. Zhang, P. Zhu, P. Chen, X. Tang, C. Li, and L. Jiao, "Foreground refinement network for rotated object detection in remote sensing images," *IEEE T GEOSCI REMOTE.*, vol. 60, pp. 1–13, 2021.

[8] H. Ruan, W. Qian, Z. Zheng, and Y. Peng, "A Decoupled Semantic–Detail Learning Network for Remote Sensing Object Detection in Complex Backgrounds," *ELECTRONICS SWITZ.*, vol. 12, no. 14, pp. 3201, 2023.

[9] X. Li, W. Wang, W. Wu, L. Chen, X. Hu, J. Li, and J. Yang, "Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection," *Advances in Neural Information Processing Systems.*, vol. 33, pp. 21002–21012, 2020. .

[10] J. Wang, C. Xu, W. Yang, and L. Yu, "A normalized Gaussian Wasserstein distance for tiny object detection," 2021, *arXiv:2110.13389*.

[11] J. Chen, S. Kao, H. He, W. Zhuo, S. Wen, C. H. Lee, and S. H. Chan, "Run, Don't Walk: Chasing Higher FLOPS for Faster Neural Networks," 2023, *arXiv:2303.03667*.

[12] Y. Liu, H. Li, C. Hu, S. Luo, Y. Luo, and C. Chen, "Learning to Aggregate Multi-Scale Context for Instance Segmentation in Remote Sensing Images," 2022, *arXiv:2111.11057*.