Annals of Emerging Technologies in Computing (AETiC)
Vol. 9, No. 2, 2025

Research Article

Efficient Object Detection in Remote Sensing Images Using Quantitative Augmentation and Competitive Learning

Huaxiang Song[1,*], Junping Xie[2], Yan Zhang[3], Yang Zhou[1], Wenhui Wang[1], YingYing Duan[1] and Xinyi Xie[1]

[1]Hunan University of Arts and Science, China
cn11028719@huas.edu.cn; zhouyang@huas.edu.cn; 2412@huas.edu.cn; 666888spring@st.huas.edu.cn; xiliyuanyu@st.huas.edu.cn
[2]Kunming University of Science and Technology, China
hnxiejunping@163.com
[3]Huanggang Normal University, China
89996457@qq.com
*Correspondence: cn11028719@huas.edu.cn

Received: 6th November 2024; Accepted: 23rd March 2025; Published: 1stApril 2025

Abstract: Object detection in remote sensing images (RSIs) is crucial in Earth observation. However, current approaches often overlook key characteristics of RSIs, resulting in models that fail to balance accuracy and computational efficiency. To the authors' knowledge, these limitations stem from the inherent scarcity and complexity of RSI samples, which cannot be fully resolved by solely modifying the model architecture. To address these challenges, we propose QACL-Net, an object detection method built on the Faster R-CNN framework, which significantly enhances the performance of CNN-based detectors for RSI recognition while maintaining fast inference speeds. QACL-Net incorporates several innovative techniques. Firstly, we introduce the quantitative augmentation (QA) strategy to address RSI sample scarcity. Secondly, we propose the equal-quadrate mosaic (EQM) algorithm to improve the effectiveness of the traditional mosaic technique for RSI detection. Thirdly, we implement the competitive learning (CL) strategy to resolve the problem of redundant feature fusion in the feature pyramid network. Crucially, the proposed enhancement techniques are integrated into three plug-and-play modules. To evaluate the proposed method, we develop two variants of QACL-Net by utilizing an EfficientNet-B0 and EfficientNet-B3 backbone model for the detector architecture, respectively. Extensive experiments on two widely used RSI datasets demonstrate that QACL-Net outperforms 31 advanced methods since 2022 on the DIOR20 dataset. Specifically, QACL-Net-B3 achieves a 6.9% improvement in accuracy on the challenging DIOR20 dataset. Additionally, QACL-Net-B3 reduces model size by 33% and increases inference speed by 17% compared to the baseline model. In summary, our work highlights the significant impact of RSI sample scarcity, noisy backgrounds, and feature fusion redundancy on object detection performance. Theoretically, our approach can be seamlessly integrated with other detection models, as the QA, EQM, and CL modules require only minimal modifications to the model structure.

Keywords: ***Competitive Learning; Equal-Quadrate Mosaic; QACL-Net; Quantitative Augmentation; Remote Sensing Object Detection***

1. Introduction

Remote sensing images (RSIs) have become indispensable across various fields, offering unique advantages in areas such as Earth observation [1], environmental monitoring [2], agriculture [3], and aquaculture [4]. Object detection methods in RSIs play a crucial role by automating the identification of objects as well as determining their location and quantity [5–6]. Currently, deep learning approaches are

Huaxiang Song, Junping Xie, Yan Zhang, Yang Zhou, Wenhui Wang *et al.*, "Efficient Object Detection in Remote Sensing Images Using Quantitative Augmentation and Competitive Learning", *Annals of Emerging Technologies in Computing (AETiC)*, Print ISSN: 2516-0281, Online ISSN: 2516-029X, Vol. 9, No. 2, 1st April 2025, pp. 1-19, Published by International Association for Educators and Researchers (IAER), DOI: 10.33166/AETiC.2025.02.001, Available: http://aetic.theiaer.org/archive/v9/v9n2/p1.html.

widely adopted for object detection due to their ability to manage the complexity and diversity of data found in RSIs [7-8].

Object detection in RSIs generally relies on frameworks originally developed for natural images. These frameworks are typically categorized into two types based on their detection algorithms: one-stage and two-stage methods. One-stage methods, such as You Only Look Once (YOLO), predict object classes and bounding boxes simultaneously, prioritizing speed and making them well-suited for real-time applications [9]. In contrast, two-stage algorithms, like Faster R-CNN, first generate region proposals (i.e., likely object locations) and then classify these regions and refine the bounding boxes, achieving higher accuracy but with slower processing times [10]. Both types employ a backbone model for feature extraction and a detector neck for feature refinement, aiding in the prediction of object categories and locations. However, unlike natural images, RSIs contain numerous small, densely packed objects within complex backgrounds, which challenges the detection effectiveness in background differentiation and small-object precision. Consequently, researchers have proposed various optimization approaches to improve the accuracy of these frameworks in RSI detection.

YOLO models often focus on enhancing speed and reducing complexity [11]. Techniques include lightweight backbone models and efficient multi-scale processing modifications, which improve detection accuracy without compromising processing time [12]. On the other hand, Faster R-CNN optimizes target precision through adjustments in region proposal networks [13] and the introduction of attention mechanisms [14], which refine object localization in complex or high-density settings. Both frameworks benefit from adaptive loss functions and spatial pooling [15-16], which improve detection speed and accuracy under challenging remote sensing samples. However, RSIs are characterized by noisy backgrounds and multi-scale objects [17]. Additionally, the scarcity of RSI samples is common due to the specialized focus of remote sensing research [18]. Therefore, RSI recognition often requires tailored training strategies distinct from those used for natural images, ensuring superior model performance [19–20].

To address the unique challenges in RSIs, researchers have developed various optimization strategies for the Feature Pyramid Network (FPN), emphasizing its key capabilities in feature fusion [21]. Many approaches focus on multi-scale feature selection, using adaptive layers to enhance the detector's ability to identify small and large objects across diverse scales [22]. Alternatively, some methods employ gating functions to filter out irrelevant information, thereby improving the model's accuracy by emphasizing crucial features [23]. Additionally, related studies suggest that multi-stage integration modules and adaptive fusion techniques effectively blend information from different layers, managing complex image textures and varying object scales [24]. Similarly, attention-based mechanisms are embedded within FPNs to refine spatial relationships [25], further improving object localization, particularly for targets with irregular shapes and orientations. However, some inherent characteristics of RSIs are due to the data distribution, like variability and scarcity across samples. Therefore, optimization strategies focusing solely on model structure are insufficient to overcome the challenges in RSI recognition [26–27].

To address the challenges of variability and scarcity in RSIs, researchers have introduced various data augmentation (DA) strategies to enhance model robustness by increasing dataset diversity and mitigating overfitting [28]. Studies indicate that spatial transformations, such as rotation, flipping, and cropping, effectively simulate diverse object orientations, improving model generalization, particularly for small or uniquely shaped objects [29]. Likewise, techniques like MixUp and Mosaic combine multiple image samples to create synthetic images, introducing novel contexts and complex backgrounds that enhance model adaptability to varying scene compositions [30]. Additionally, noise perturbation and color adjustments replicate environmental changes, such as lighting and atmospheric variations, further enhancing model resilience across diverse conditions [31]. Synthetic data generation through generative adversarial networks also supports augmentation in data-scarce scenarios by expanding dataset volume without incurring the costs of manual labeling [32]. However, to the authors' knowledge, previous studies still failed to address several challenges inherent to RSI samples, leading to suboptimal model performance.

Remote sensing imagery often suffers from low quality due to varying imaging conditions and processes. As a result, the RSI samples shown in Figure 1 exhibit inconsistent image quality, with only the left samples in each category meeting acceptable standards. This inherent challenge complicates the

differentiation between inter-class and intra-class samples. To develop robust models, current DA strategies typically degrade high-quality samples to simulate low-quality ones, aiming to replicate the diverse imaging conditions encountered in remote sensing. However, these qualitative DA methods, which are typically used for natural images, apply this transformation to all training samples throughout the training process.

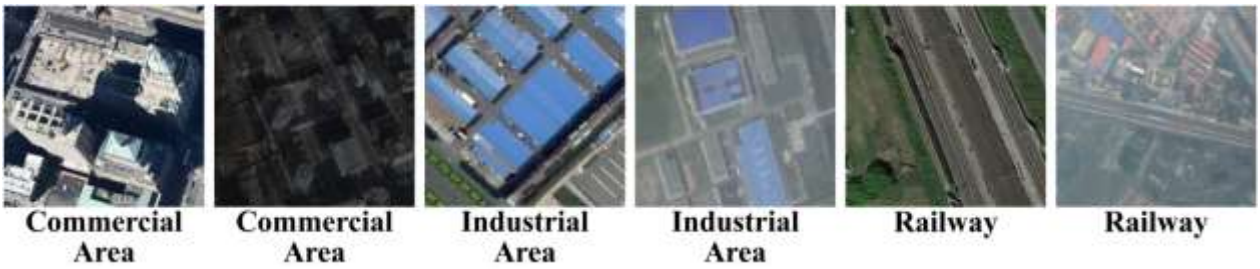

Figure 1. Imaging Quality Variations in RSIs

In contrast, real-world RSI samples are usually pre-screened by algorithms before being released, ensuring that most of the dataset maintains acceptable image quality. As a result, applying these qualitative DA techniques to RSI detection introduces a significant discrepancy between the training data and real-world applications, where the majority of samples retain good imaging quality. This misalignment leads to suboptimal performance of RSI detectors, as the training data distribution created by these DA methods differs from that of real-world data.

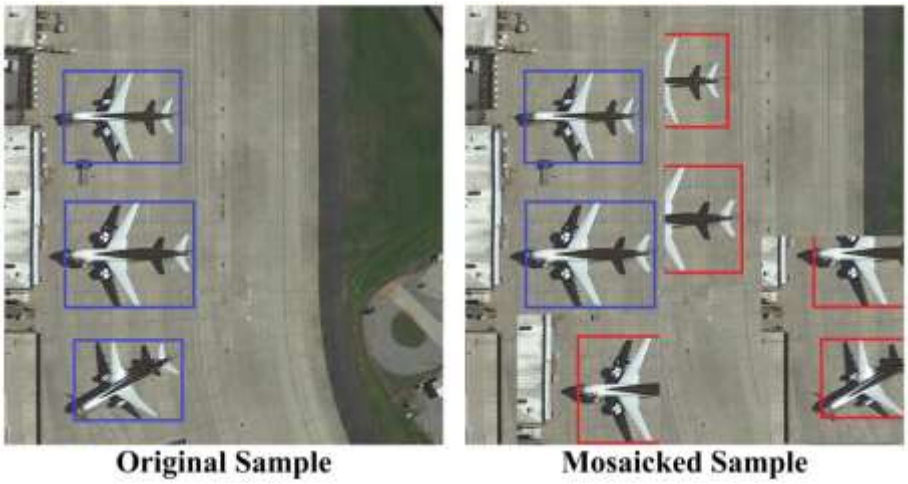

Figure 2. Fragmented Object Representations in RSIs through Mosaic DA

DA techniques such as MixUp and Mosaic are effective for natural images, which typically contain larger objects. However, remote sensing imagery often comprises numerous smaller objects of varying sizes within a scene. Figure 2 illustrates how the Mosaic technique generates synthetic images by randomly selecting four patches from training samples. This process often leads to fragmented representations of objects because RSI samples typically include many fine-grained elements. Consequently, models trained on these fragmented representations struggle to accurately identify and locate target objects, thereby increasing the likelihood of prediction errors in object detection.

Similarly, the MixUp technique faces challenges when applied to RSI samples. MixUp generates synthetic samples by linearly interpolating between two images, blending their pixel values. However, this process poses difficulties for RSIs, as blending pixel values from different RSI samples can result in unnatural transitions between objects, leading to unrealistic representations that do not reflect real-world scenarios. This synthetic algorithm confuses the model during training, as the generated samples fail to accurately represent the true variation in object appearances or spatial relationships typical in RSIs. Consequently, training with such mixed samples may degrade the model's ability to accurately detect and classify objects in remote sensing tasks.

Multi-layer feature fusion enhances a model's capacity to detect objects of varying sizes. However, the complex backgrounds in RSI samples often introduce substantial redundant information during the fusion process. As illustrated in Figure 3, when feature fusion includes a road background with two

vehicles (highlighted by blue rectangles), it provides useful context for the detector to identify the vehicles. In contrast, a grassland background (depicted by a red rectangle) complicates detection, as vehicles are more frequently seen on roads than on grasslands. Current FPN-based methods have not fully addressed this challenge, leading to ineffective feature fusion with unnecessary background information, which can degrade the model's performance.

Figure 3. The Impact of Different Backgrounds on Feature Fusion

In this paper, we introduce QACL-Net, a Faster R-CNN-based detection framework designed to address key challenges in RSI detection by integrating several innovative techniques. First, we propose a quantitative augmentation (QA) strategy that regulates the frequency of image transformations in a flexible framework to address limitations in existing augmentation approaches. Second, we develop an equal-quadrate mosaic (EQM) algorithm to generate synthetic training samples while maintaining the integrity of object shapes within RSIs. Third, we implement a competitive learning (CL) strategy for feature fusion in FPNs, achieving notable improvements in detection accuracy with minimal impact on inference speed. Additionally, we replace the ResNet-50 backbone in the Faster R-CNN model with the EfficientNet, optimizing the model's size and efficiency. Specifically, we incorporate the QA strategy, EQM algorithm, and CL strategy as three plug-and-play modules, referred to as the QA module, EQM module, and CL module, respectively, to enhance the proposed method's adaptability and transferability.

We evaluate the performance of the proposed QACL-Net on two RSI datasets. The experimental results demonstrate that QACL-Net outperforms 39 methods since 2022 on the challenging DIOR20 dataset. Notably, QACL-Net achieves a 6.9% improvement in accuracy on the DIOR20 dataset, significantly surpassing other top-performing models. Furthermore, QACL-Net achieves a 33% reduction in model size and a 17% increase in inference speed compared to the baseline model. The key contributions of this paper are as follows:

(1) We introduce the QA strategy and the EQM algorithm to enhance the object detection performance of the Faster R-CNN model in RSIs. These novel techniques effectively address inherent challenges in RSIs, leading to significant improvements in detection accuracy.

(2) We propose the CL strategy to mitigate the issue of redundant feature fusion in current detection frameworks. This strategy results in substantial improvements in object detection accuracy in RSIs with minimal computational overhead.

(3) We implement the QA strategy, EQM algorithm, and CL strategy as plug-and-play modules, offering the potential for integration with other existing detection methods. Additionally, the QACL-Net-B3 model demonstrates superior performance in terms of both accuracy and inference speed, providing a practical solution for object detection in RSIs.

2. Related Works

2.1. Strategies for Optimizing Detector Frameworks

Since 2022, advances in RSI object detection have highlighted the challenges posed by noisy backgrounds and varying object scales. To address these issues, MFICDet [11] employs a dual strategy approach. It introduces a positive and negative feature guidance module to suppress background noise and a global feature information complementary module to enhance detection in complex environments. ProEnDet [13] introduces an anchor-free detector that utilizes the weighted bidirectional FPN and probability enhancement techniques to improve the distinction between foreground and background objects. HA-MHGEN [14] stands out with its hybrid attention-driven, multi-stream hierarchical graph network. It captures both spatial and semantic relationships using self-attention mechanisms and graph embedding. SHDet [15] enhances feature representation across spatial hierarchies by incorporating a spatial hierarchy perception component and applying hard sample metric learning, thereby reducing intra-class variability and boosting detection performance. MOD-Net [16] tackles scale variations and background complexity by integrating multi-receptive field features and relation-connected attention into the Faster R-CNN framework, resulting in more robust detection. Collectively, these methods demonstrate substantial performance gains over their baseline models. Nevertheless, their optimization strategies are still one-dimensional, exposing inherent limitations.

Researchers have proposed several strategies to enhance their detector frameworks. For example, MCFCE-Net [33] integrates a multi-scale contextual feature enhancement module, employing recursive convolution and attention mechanisms to capture scale-dependent information and reduce background noise. TBNet [34] introduces two specialized modules to address the detection of small and weak objects. Its texture-aware module captures texture details through pixel correlations, while the boundary-aware fusion module highlights object edges to improve spatial localization. In contrast, TRD [35] combines CNNs with a modified Transformer architecture to leverage both local and global feature representations. By using a multi-layer Transformer, TRD addresses the limitations of CNNs in handling long-range dependencies by aggregating global spatial features. SAENet [36] targets weakly supervised object detection through self-supervised and adversarial learning techniques. Its adversarial dropout activation block dynamically obscures discriminative object parts to highlight more comprehensive instance features. GLC-Net [37] employs a dual attention mechanism in an end-to-end architecture for multi-size object detection. Its MobileNet backbone extracts multi-layer features, while the two-stage deep feature fusion module combines feature maps to enhance the representation of small targets. Additionally, GLNet [38] addresses the issue of varying target scales by integrating global context cues from a multi-scale perception module with local spatial correlations. SGFTHR [39] focuses on detecting small and dense objects by introducing a structure-guided feature transform module to preserve critical low-level spatial and structural information. Overall, these seven approaches primarily concentrate on model structures but lack effective measures to address the unique data distribution in RSIs.

2.2. Enhancements in FPN Optimization

Recent advancements have increasingly leveraged FPNs to address the inherent challenges posed by complex backgrounds and varying object scales in RSIs. For instance, Sw-LBPN [12] introduces a simplified bidirectional FPN to facilitate multi-scale feature aggregation. It utilizes skip connections to preserve and reuse information from small-scale objects. MSA R-CNN [21] introduces an adaptive dynamic inner lateral connection module to reduce information loss in the FPN and a distributed lightweight attention module for refined feature information processing. SIFA-Net [22] proposed two novel modules. Its adaptive feature extraction module integrates local and global features to accommodate the varying angles of small targets, and the tri-directional feature fusion module enhances the quality of feature maps through a weighted fusion mechanism. Additionally, Bayes R-CNN [23] employs a multi-level feature fusion module to minimize information loss in the FPN and a Bayesian distributed lightweight attention module to facilitate background classification for a detailed interpretation of detected objects. ABNet [24] introduces an adaptive FPN to mitigate the impact of complex backgrounds on foreground objects. It develops a context enhancement module to leverage

abundant semantic information for multi-scale object detection. Moreover, TransMIN [25] integrates cross-view feature interactions in the FPN, enhancing edge information and mitigating background interference by capturing correlations between reference features (such as spatial edge priors and channel statistics). Collectively, these approaches highlight a growing trend toward integrating multi-scale features. However, they still lack effective measures to address the efficacy of feature fusion.

Other studies have focused on the intelligent fusion of local and global information for optimizing the classic FPN. For instance, REFIPN [40] employs rotation equivariance convolution and a lightweight image pyramid module. It enhances small-scale object detection by effectively extracting features across various scales and orientations. GCF-Net [41] introduces an aware FPN for cleaner feature extraction and a group assignment strategy for more effective label distribution based on sample overlap. This approach addresses feature interference and label assignment issues in RSI samples. Similarly, RFEB-Net [42] presents a multi-scale detection framework. It incorporates a receptive field expansion block to enhance context capture and modifies the classic FPN with dilated convolutions to preserve resolution while maintaining a large receptive field. MSNet [43] introduces a partial and point-wise convolution extraction module for simultaneous spatial and channel feature extraction and a local and global information fusion module to integrate texture and semantic information. Additionally, MSNet contains a local and global information fusion pyramid that enhances small object detection by densely connecting multi-scale feature maps. MSFP-Net [44] proposes a content-aware feature upsampling and feature enhancement module for efficient feature map fusion across stages, addressing small-scale target blurring and large-scale variations in RSI samples. Lastly, GPANet [45] introduces a gated path aggregate network, which incorporates path enhancement and information filtering to optimize feature fusion by assigning importance to hierarchical convolutional layers.

Overall, these methodologies demonstrate a concerted effort to improve feature representation through sophisticated feature integration techniques. Nonetheless, the redundant feature fusion in the FPN has not been addressed.

2.3. Strategies for RSI Augmentation

Recent research has proposed various strategies to augment training datasets in RSI detection. For instance, CGDDT-Net [28] introduces a count-guided deep descriptor transforming algorithm, which automatically generates coarse object bounding boxes from class labels and per-class object counts. The TSPGC [29] approach proposes an augmentation technique called shape-prior-based generated content, addressing the limitations of RSI samples' quantity and quality. It generates shape data independently of existing training datasets and enhances its robustness through stylization. The double augmentation [30] method presents a two-step framework for ship detection in RSIs. It consists of a front augmentation step that employs a modal recognition network to mitigate differences during training and a back augmentation verification step that utilizes batch augmentation. Additionally, P2P-Net [31] implements an augmentation strategy for solid waste detection, utilizing a modified pix2pix model to improve the detail and overall quality of the generated images. MAGAN [32] integrates a framework for automatically generating content-rich synthetic images with ground-truth annotations. It involves rendering 3D CAD models to create two widely distributed synthetic aircraft image datasets and enhancing image quality through a multi-scale attention module. Moreover, MBS-Net [46] introduces an improved augmentation strategy to manage multi-scale and directional variations. It incorporates a multi-branch stacking module to capture deep target features effectively and employs a dual-channel attention mechanism to enhance discriminative feature acquisition in complex scenes. RFA-Net [47] incorporates a sequential augmentation module to enhance unlabeled data, a sparse feature reconstruction module to strengthen instance-level features for better alignment, and a pseudo-label generation module to supervise the unlabeled target domain.

Overall, these augmentation strategies mark a significant advancement in addressing the limitations associated with insufficient training data in RSI detection. However, they have not sufficiently resolved the issue of biased data distribution stemming from qualitative augmentation and often necessitate multiple steps for sample processing. Consequently, the accuracy outcomes of these methods may not be competitive.

3. Methodologies

3.1. Detection Framework

Figure 4 illustrates the detection framework of the proposed QACL-Net, which incorporates a Faster R-CNN architecture with an embedded FPN. As shown on the left side of the figure, the QA and EQM modules process RSI samples sequentially; the resulting transformed data flow is then input into the EfficientNet backbone for feature extraction. Subsequently, as indicated at the top of the figure, the FPN utilizes feature maps from the convolutional stages 2, 3, 4, and 5 of the backbone to perform feature fusion. Notably, the CL module oversees this fusion process in a competitive manner. Finally, the region proposal network generates category predictions and bounding boxes for target objects.

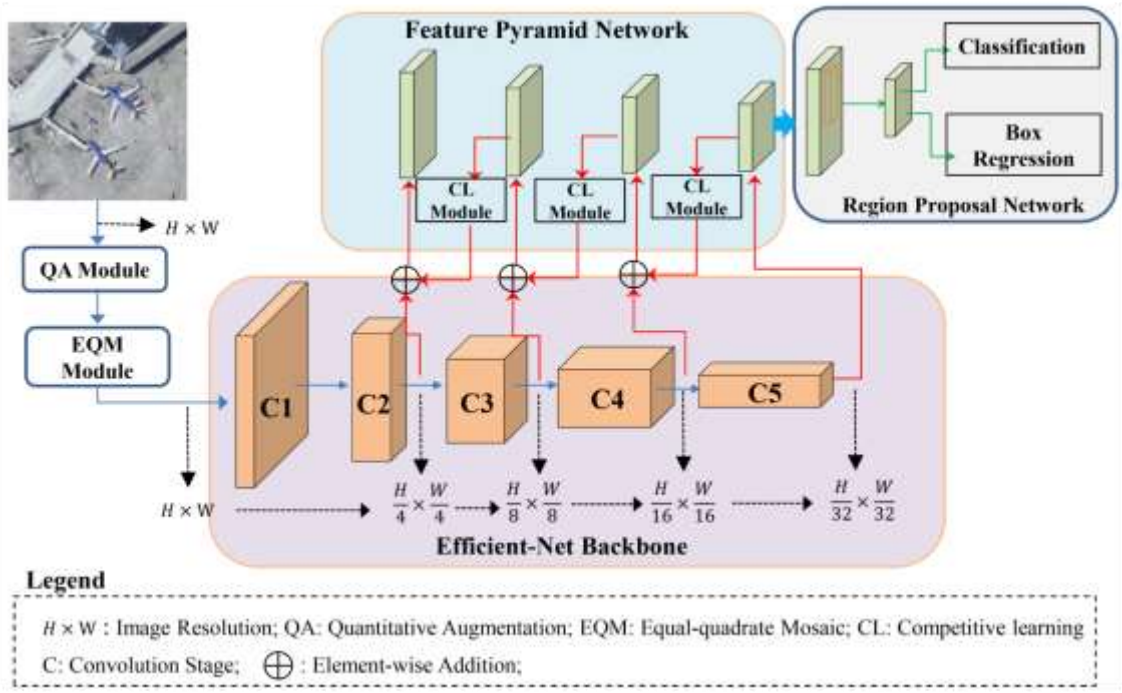

Figure 4. Detection Framework of the Proposed QACL-Net

The QA and EQM modules are active only during the training phase and are removed during inference. Furthermore, QACL-Net introduces only a few structural optimizations to the FPN. Therefore, the proposed method introduces negligible additional computational costs compared to the original Faster R-CNN framework.

3.2. Propose QA Module

Figure 5 illustrates the proposed QA module, which consists of a sequential arrangement of six QA operators. Each QA operator, as depicted at the top of Figure 5, comprises three main components: a probability sampler, a logic branch, and a transform function. For each input RSI sample, these operators either produce a transformed sample or retain the original sample based on a chance determined by the probability sampler.

Specifically, let P and Th denote the output chance from the probability sampler and the probability threshold within the logic branch, respectively. Let F and $\tilde{F}$ represent the input and output features of QA operators, while f denotes the transform function embedded within the operator. The operational mechanism of a QA operator can be described as follows:

$$\tilde{F} = \begin{cases} f(F), if\ P \le Th \\ F \quad\ , if\ P > Th \end{cases} \tag{1}$$

The six QA operators use distinct transform functions, which are color jitter, horizontal flip, vertical flip, grayscale conversion, auto contrast, and Gaussian blur. According to ablation experimental results presented in Section 4.3, the probability threshold values for the QA operators are 0.3 for the grayscale and auto contrast and 0.5 for the others. The notations of input and output features of the QA module,

along with the transform function, are consistent with those defined in Equation (1). The overall workflow of the QA module can be summarized as follows:

$$\tilde{F}_A = f_A(F) \tag{2}$$

$$\tilde{F}_B = f_B(\tilde{F}_A) \tag{3}$$

$$\tilde{F}_C = f_C(\tilde{F}_B) \tag{4}$$

$$\tilde{F}_D = f_D(\tilde{F}_C) \tag{5}$$

$$\tilde{F}_E = f_E(\tilde{F}_D) \tag{6}$$

$$\tilde{F} = f_F(\tilde{F}_E) \tag{7}$$

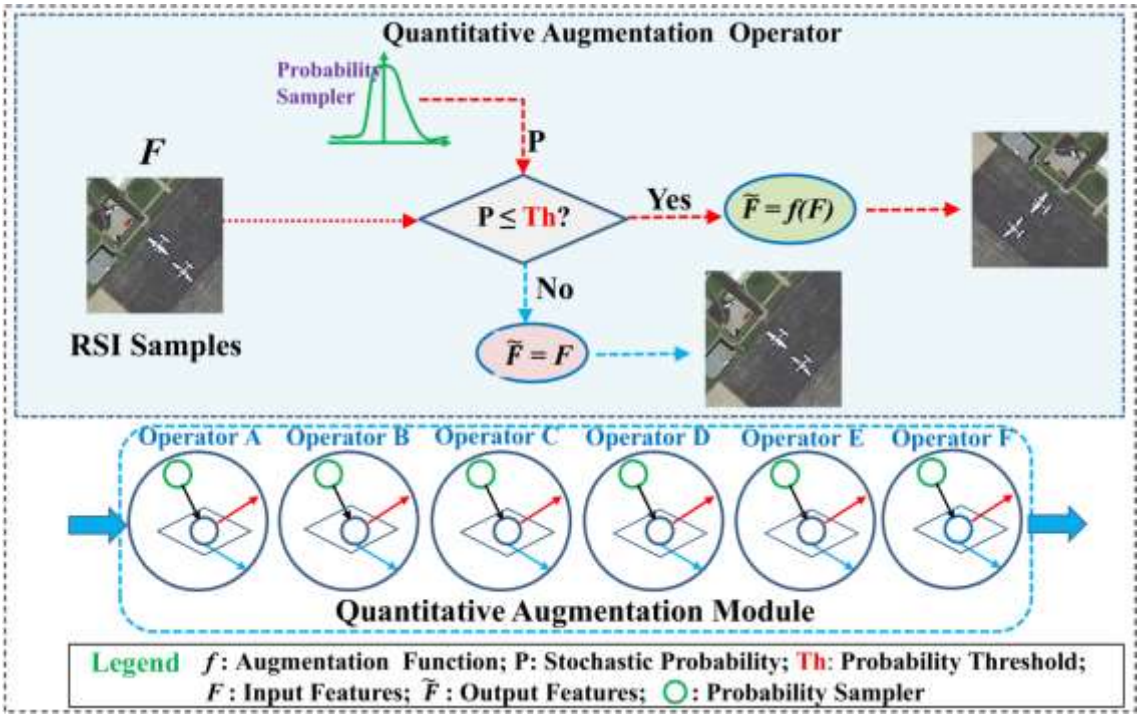

Figure 5. Architecture of the Proposed QA Module

3.3. Propose EQM Module

Algorithm 1 presents the pseudocode for the proposed EQM. In detail, line 3 initializes a variable for a 2× image, which has dimensions twice the height and width of the current input RSI sample. In line 4, three images are randomly selected from the sample batch, along with their corresponding bounding boxes. Line 5 copies the features from the current sample and the three selected images into the 2× image variable. Subsequently, lines 6 and 7 resize the 2× image variable to match the original dimensions of the input RSI sample, adjusting the bounding boxes accordingly. Finally, lines 8 to 10 generate synthetic mosaic samples, ensuring that the bounding boxes are updated to reflect the new images.

Algorithm 1. Procedures of the proposed EQM

Input: A batch of RSI samples, denoted as x_i, with corresponding bounding boxes b_i. The height and width of the RSI sample are represented as H and W, respectively.

Output: A batch of transformed RSI samples, denoted as $\tilde{x}_i$, with corresponding bounding boxes $\tilde{b}_i$.

1: Initialize an empty list, denoted as xList.

2: **For** iteration = 1 to $length(x_i)$ **Do**

3: Initialize a temp image variable, denoted as x_T, with its channel number, height, and width at: 3, H×2, and W×2, respectively.
Initialize an empty bounding box list, denoted as bList.

4: Randomly sampling three images from the sample batch, denoted as x_A, x_B, x_C, respectively.
Obtain the bounding boxes of x_A, x_B, x_C, denoted as b_A, b_B, b_C, respectively.

5: x_T[:, 0:H, 0:W] = $x_{iteration}$[:, :, :]
x_T[:, 0:H, W:2W] = x_A[:, :, :]
x_T[:, H:2H, 0:W] = x_B[:, :, :]
x_T[:, H:2H, W:2W] = x_C[:, :, :]
Fix the coordinates in b_A, b_B, and b_C accordingly.

6: Resize the height and width of x_T to H and W.
Fix the coordinates in x_T accordingly.
7: Add x_T into xList.
Add $b_{iteration}$, b_A, b_B, and b_C into bList.
8: **End For**
9: $\tilde{x}_i$ = xLis, $\tilde{b}_i$ = bList
10: **Return** $\tilde{x}_i$ and $\tilde{b}_i$

Additionally, the EQM module incorporates a probability sampler that regulates the frequency of mosaic sample generation, following the method outlined in Equation (1). The probability threshold for the EQM module is set to 0.1, as higher probability values during training lead to suboptimal detection accuracy.

3.4. Propose CL Module

The CL structure is inspired by the human cognitive model, wherein two distinct teams work collaboratively to solve a common task, utilizing competition to enhance the final solution. Figure 6 illustrates the architecture of the proposed CL module, which consists of two main components. The feature fusion module within the FPN, depicted on the left side of Figure 6, integrates the CL module as its initial layer, followed by two additional convolutional blocks. In contrast, the feature fusion layer in a standard FPN architecture only includes the two convolutional blocks shown in Figure 6.

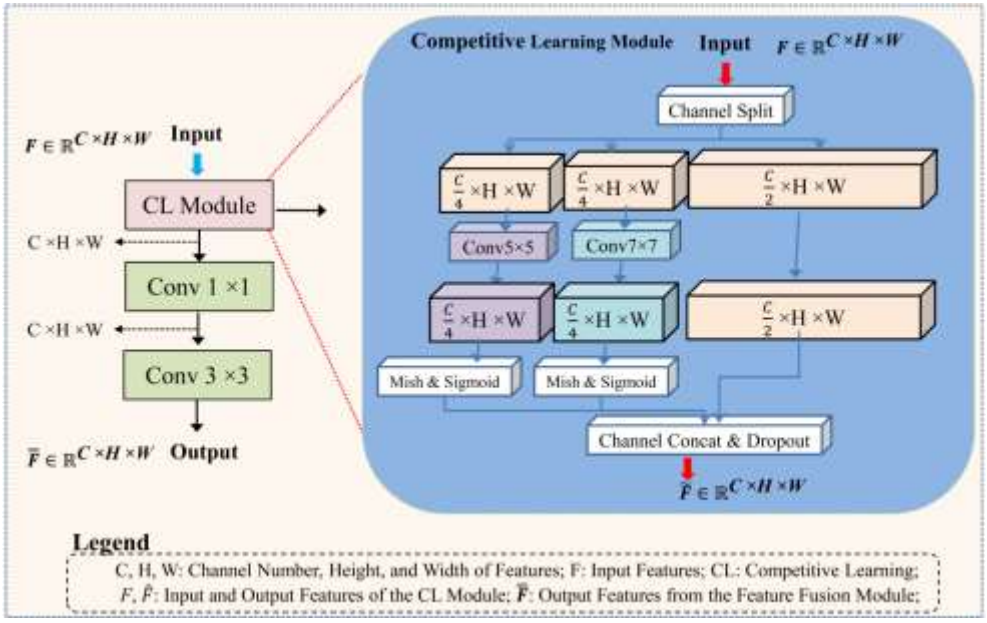

Figure 6. Structure of the Proposed CL Module

The CL module, detailed on the right side of Figure 6, introduces an innovative structure. Specifically, it divides the input features along the channel dimension into four groups of equal size. The first and second groups are processed separately by 5×5 and 7×7 convolutional layers, respectively, followed by activation functions. In contrast, the third and fourth groups remain unaltered. Finally, the split features are recombined along the channel dimension, preserving the original number of channels in the output features as in the input to the CL module.

3.5. Backbone Models

We assess the effectiveness of the proposed QACL-Net by developing two detectors, each utilizing a different EfficientNet variant as the backbone: EfficientNet-B0 and EfficientNet-B3. EfficientNets, as described in [48], are well-known CNN models that incorporate channel attention mechanisms. In general, EfficientNet-B0 and EfficientNet-B3 demonstrate superior performance compared to ResNet-50, especially in handling RSI classification tasks while maintaining compact model sizes.

3.6. Dataset and Division

We employed two datasets, NWPU10 and DIOR20 [7], to evaluate the performance of the proposed QACL-Net. Table 1 provides a detailed summary of the dataset characteristics and the training ratio (TR) configurations. NWPU10 is a relatively small-scale dataset, and previous studies have often used a TR exceeding 50% to maintain accuracy. On the other hand, DIOR20 is considerably more challenging due to its larger size. DIOR20 has been pre-divided into training, testing, and validation subsets since its release, with proportions of 25%, 25%, and 50%, respectively. For our experiments, we use the original training subset, corresponding to a 25% TR of DIOR20, for model training. We then evaluate and report the performance using the original validation subset, which represents 50% of the DIOR20 for testing.

Table 1. Summary of Detection Dataset and Training Ratio Settings

Dataset	Category Number	Total Images	Image Size	Total Instances	TRs	Testing Ratios
NWPU10	10	800	500~1100 × 500~1100 (varied)	3,775	50%	50%
DIOR20	20	23,463	800 × 800 (fixed)	192,472	25%	50%

3.7. Performance Evaluation Metrics

We used mean average precision (mAP) as the evaluation metric for detection performance. In this context, 'TP' denotes the number of correctly identified positive instances, 'FP' refers to the number of negative instances mistakenly classified as positive by the model, and 'FN' indicates the number of positive instances incorrectly classified as negative. Precision (P) and recall (R) are defined as follows:

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

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

IoU is a measure of the overlap between the ground truth bounding boxes (represented as B_T) and the predicted bounding boxes (represented as B_P). IoU can be expressed as follows:

$$IoU = \frac{B_P \cap B_T}{B_P \cup B_T} \tag{10}$$

Average Precision (AP) is defined as the area under the precision-recall curve, as illustrated below:

$$AP = \int_0^1 P(R) \cdot d(R) \tag{11}$$

All mAP results presented in this study correspond to the mean AP across categories, calculated at an IoU threshold of 0.5.

3.8. Implementation Details

The experiments were conducted using four Nvidia 4070Ti-Super GPUs with PyTorch version 2.10.0 in an Ubuntu 20.04 environment. Training was performed over 72 epochs using the Adam-W optimizer with a weight decay of 10^{-6}. The input resolution was set to 800^2 for the DIOR20 dataset and 1200^2 for the NWPU10 dataset. The training batch size was 32, and the initial learning rate, regulated by a cosine decay schedule, was set to 0.00005. The reported results represent the best outcomes from three independent trials.

4. Experimental Results

We compared the performance of the proposed QACL-Net with 31 studies published since 2022. However, many of these studies did not demonstrate competitive performance. To conserve space, we have included performance comparison data for only the 13 highest-ranked methods in the literature. Additionally, many of the compared methods reported their performance on either the NWPU10 or DIOR20 dataset separately. Therefore, the total number of compared methods on each dataset is slightly less than 13.

Tables 2 and 3 present the mAP comparison results for the NWPU10 and DIOR20 datasets, respectively. In these tables, the "roadmap" column lists various detection frameworks, and the symbol "–" indicates instances where the relevant literature does not disclose data. Values in bold represent superior performance within a given column.

4.1. Accuracy Result on NWPU10

Table 2. Comparison of mAP (%) and AP (%) per Category on NWPU10

Method	Roadmap	mAP	C01	C02	C03	C04	C05	C06	C07	C08	C09	C10
MFICDet [11]	YOLO	96.4	99.9	95.7	69.8	**99.3**	93.3	76.0	97.9	84.2	90.2	**97.7**
Sw-LBPN [12]		93.8	99.5	95.2	95.2	97.8	92.1	96.4	99.5	93.6	82.9	85.9
TBNet [34]		95.4	87.5	43.3	100.0	86.6	**100.0**	92.1	95.9	**99.7**	85.2	92.0
HA-MHGEN [14]	Faster R-CNN	93.4	97.2	88.9	98.7	83.1	94.2	99.0	90.5	87.6	**97.7**	97.1
ABNet [24]		92.3	92.5	**97.7**	97.7	99.2	95.9	98.8	94.2	69.0	96.6	94.2
GLNet [38]		91.8	100.0	84.4	98.5	81.6	88.2	**100.0**	97.2	88.4	90.9	88.7
MSNet [43]	Self-designed Framework	96.0	**100.0**	88.9	96.9	99.9	96.8	98.2	97.6	97.1	88.7	95.8
CGDDT-Net [28]	Faster R-CNN With Tailored-DA	93.2	88.2	91.1	90.5	89.2	91.9	91.1	91.4	94.0	93.2	88.2
QACL-Net-B0	Ours	94.2	99.7	83.4	99.6	97.9	93.2	89.2	97.6	74.3	58.3	85.1
QACL-Net-B3		**95.1**	99.0	90.4	**100.0**	98.7	96.6	97.1	**100.0**	84.7	78.4	92.5

*C01 to C10 represent the categories: airplane, ship, storage tank, baseball diamond, tennis court, basketball court, ground track field, harbor, bridge, and vehicle, respectively.

Table 2 presents a comparison of the mAP and AP results per category for various methods evaluated on the NWPU10 dataset. Among the methods, MFICDet achieves the highest mAP of 96.4%, excelling particularly in categories such as airplane (99.9%), ship (95.7%), and vehicle (97.7%).

Notably, MFICDet, MSNet, and TBNet were trained using a training ratio significantly exceeding 60%, which, combined with a testing ratio of less than 25%, leads to performance saturation due to the limited size of the NWPU10 dataset. As a result, these models, while achieving high mAP values, are more susceptible to overfitting due to the sparse number of testing samples. In contrast, the proposed QACL-Net-B0 and QACL-Net-B3 models were trained with a more balanced training-to-testing ratio of 50%, which includes a larger proportion of testing samples and mitigates the risk of overfitting.

As a result, QACL-Nets demonstrate a more robust and competitive performance across categories, achieving an mAP of 94.2% for QACL-Net-B0 and 95.1% for QACL-Net-B3, with consistent and strong results in categories such as airplane (99.7%) and storage tank (100%). These results highlight the effectiveness of the QACL-Nets in handling the challenges posed by the small-scale nature of the NWPU10 dataset, ensuring improved generalization without overfitting.

4.2. Accuracy Result on DIOR20

Table 3 compares the performance of various models on the DIOR20 dataset. QACL-Net-B0 achieves an mAP value of 75.2%, while QACL-Net-B3 surpasses it with an mAP of 77.6%, significantly outperforming all the other models like MFICDet (72.0%) and TBNet (73.6%).

QACL-Net-B0 and QACL-Net-B3 demonstrate strong category-specific performance, particularly in categories like C11 (tennis court), C14 (expressway service area), and C16 (airplane), where both models maintain competitive accuracy. For instance, QACL-Net-B3 delivers excellent results in C14 (expressway service area) with 91.3% AP and C19 (windmill) with 89.5% AP, outperforming other models such as Sw-LBPN (90.0% in C14) and SIFA-Net (72.9% in C19). Additionally, QACL-Net-B0 and QACL-Net-B3 show superior robustness in categories such as C09 (airplane) and C10 (ground track field), where they achieve solid AP scores of 84.8% and 85.3%, respectively. Despite their strong performance, QACL-Nets face some challenges in categories like C03 (vehicle) and C08 (harbor), where other models like MSNet and TSPGC demonstrate better performance, particularly due to optimization strategies for certain specialized objects. Overall, QACL-Net-B3 delivers the most balanced performance across all categories, proving to be highly effective for remote sensing image detection tasks.

The results from Table 3 indicate that QACL-Nets significantly improve detection accuracy compared to existing methods. They maintain competitive performance across diverse categories in the DIOR20 dataset, particularly excelling in handling challenging tasks involving various samples.

Table 3. Comparison of mAP (%) and AP (%) per Category on DIOR20

Method	Roadmap	mAP	C01	C02	C03	C04	C05	C06	C07	C08	C09	C10
MFICDet [11]	YOLO	72.0	54.1	71.4	63.3	81.0	42.6	72.5	57.5	68.7	62.1	73.1

Sw-LBPN [12]		73.9	82.8	84.9	75.0	89.8	47.8	78.0	69.0	68.9	65.5	81.5
SIFA-Net [22]		75.4	90.1	76.8	92.8	83.5	45.9	90.3	65.6	66.8	59.9	76.8
TBNet [34]		73.6	64.9	**86.8**	76.6	89.2	50.6	80.0	74.3	86.4	74.7	82.5
HA-MHGEN [14]	Faster R-CNN	74.7	88.9	77.1	52.3	81.5	**87.2**	78.1	**89.5**	**92.1**	72.2	71.4
MSA R-CNN [21]		74.3	92.9	73.8	93.2	87.3	43.0	**90.6**	58.9	69.2	58.0	83.3
Bayes R-CNN [23]		74.6	**93.6**	73.5	**93.5**	87.4	47.2	89.9	59.0	68.1	59.2	83.2
ABNet [24]		72.8	66.8	85.0	74.9	87.7	50.3	78.2	67.8	85.9	74.2	79.7
GLNet [38]		70.7	62.9	83.2	75.3	72.0	50.5	67.4	79.3	51.8	62.6	43.4
GCF-Net [41]		73.3	62.8	86.5	74.8	89.2	49.2	76.6	72.5	85.7	75.1	81.3
MSNet [43]	Self-designed Framework	75.3	95.3	67.4	91.0	**90.2**	44.6	82.6	49.2	78.4	64.8	73.4
TSPGC [29]	Faster R-CNN With Tailored-DA	76.3	-	-	-	-	-	-	-	-	-	-
QACL-Net-B0	Ours	75.2	82.2	71.8	53.3	67.1	79.1	69.9	76.1	52.9	74.1	84.8
QACL-Net-B3		**77.6**	85.2	78.9	55.2	71.9	82.0	69.4	77.0	58.7	**79.9**	**85.3**
Method	Roadmap	mAP	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20
MFICDet [11]	YOLO	72.0	76.5	42.8	56.0	71.8	57.0	63.5	81.2	53.0	43.1	80.9
Sw-LBPN [12]		73.9	78.6	62.8	61.5	90.0	74.8	75.9	87.3	67.1	56.9	79.2
SIFA-Net [22]		75.4	80.9	57.6	65.2	79.5	91.1	81.9	89.8	62.7	72.9	78.0
TBNet [34]		73.6	83.6	56.4	64.9	79.1	77.8	58.1	88.2	69.4	44.6	87.0
HA-MHGEN [14]	Faster R-CNN	74.7	85.2	74.2	75.3	71.2	46.9	86.8	69.2	52.5	71.9	71.0
MSA R-CNN [21]		74.3	84.2	57.3	62.4	68.8	**91.8**	81.3	**90.9**	53.7	72.2	74.5
Bayes R-CNN [23]		74.6	83.9	57.6	62.2	68.4	92.4	81.5	90.7	55.8	71.5	73.4
ABNet [24]		72.8	81.2	55.4	61.6	75.1	74.0	66.7	87.0	62.2	53.6	89.1
GLNet [38]		70.7	83.0	**86.2**	70.9	81.1	72.0	53.7	81.3	65.5	81.8	89.2
GCF-Net [41]		73.3	83.8	60.2	62.7	72.7	77.3	61.9	88.0	**69.9**	47.0	**89.7**
MSNet [43]	Self-designed Framework	75.3	85.1	66.4	61.8	94.4	90.4	84.5	94.8	48.1	58.9	85.4
TSPGC [29]	Faster R-CNN With Tailored-DA	76.3	-	-	-	-	-	-	-	-	-	-
QACL-Net-B0	Ours	75.2	89.8	72.2	89.4	90.3	83.4	88.4	81.5	48.6	88.9	60.3
QACL-Net-B3		**77.6**	**91.0**	73.4	**90.4**	**91.3**	84.3	**91.2**	81.2	52.1	**89.5**	63.4

*C01 to C20 represent the categories: golf field, expressway toll station, vehicle, train station, chimney, storage tank, ship, harbor, airplane, ground track field, tennis court, dam, basketball court, expressway service area, stadium, airport, baseball field, bridge, windmill, and overpass, respectively.

4.3. Ablation Experiments

In this section, we present three sets of ablation experiments conducted to evaluate the performance of the proposed QACL-Net. The results of these experiments are displayed in Tables 4, 5, and 6, respectively.

4.3.1. Ablation Experiments on the Proposed Modules

Table 4 presents a comparison of mAP results for QACL-Net-B3, illustrating the impact of sequentially embedding the QA, EQM, and CL modules into the Faster R-CNN framework. The Faster R-CNN-R50-FPN is the baseline model, representing a standard Faster R-CNN detector with a ResNet50 backbone. Increases in mAP compared to this baseline are highlighted in blue.

Table 4. Impact of Modules on mAP (%) Performance of QACL-Net-B3

Model	QA module	EQM module	CL module	DIOR20
Faster R-CNN-R50-FPN (Baseline)	×	×	×	57.8
QACL-Net-B3 (Ours)	×	×	×	68.4 ↑ 10.6
	✓	×	×	73.3 ↑ 15.5
	✓	✓	×	75.6 ↑ 17.8
	✓	✓	✓	77.6 ↑ 19.8

Table 4 presents the performance of QACL-Net-B3 on the DIOR20 dataset, highlighting the impact of the QA, EQM, and CL modules on mAP. The baseline model, which does not incorporate any of these modules, achieves an mAP of 57.8%. The QACL-Net-B3 model, which does not incorporate any of these modules, achieves an mAP of 68.4%. When the QA module is activated, the mAP increases to 73.3%, representing an improvement of 15.5%. Further enhancements are observed when both the QA and EQM

modules are employed, resulting in an mAP of 75.6%, which marks a 17.8% increase. Finally, activating all three modules—QA, EQM, and CL—led to the highest mAP of 77.6%, indicating an overall improvement of 19.8% over the baseline. These results clearly demonstrate the substantial contributions of each module to the performance of QACL-Net-B3, emphasizing the effectiveness of the integrated framework.

4.3.2. Ablation Experiments on DA Techniques

In Table 5, we present an ablation study that evaluates the effect of classic DA techniques on the performance of the QACL-Net-B3 detector when tested on the NWPU10 dataset. This study investigates how the model's performance degrades when different DA methods are consistently applied during training, while the remaining DA techniques follow the settings as we have proposed. Specifically, T01 to T06 correspond to the following augmentation strategies: color jitter (T01), horizontal flip (T02), vertical flip (T03), grayscale conversion (T04), auto contrast (T05), and Gaussian blur (T06).

Table 5. Impact of DA techniques on mAP (%) Performance of QACL-Net-B3

Model	Ablation Components						NWPU10
	T01	T02	T03	T04	T05	T06	mAP(%)
QACL-Net-B3 (Baseline)	-	-	-	-	-	-	95.1
QACL-Net-B3 (Training by Classic DA Techniques)	✓	-	-	-	-	-	94.6 ↓0.5
	-	✓	-	-	-	-	94.6 ↓0.5
	-	-	✓	-	-	-	93.9 ↓1.2
	-	-	-	✓	-	-	81.5 ↓13.6
	-	-	-	-	✓	-	94.8 ↓0.3
	-	-	-	-	-	✓	94.8 ↓0.2

As shown in Table 5, training with classic data augmentation (DA) techniques results in significant performance degradation across most configurations. For instance, applying color jitter (T01) or horizontal flip (T02) individually leads to a decrease of 0.5%, yielding a mean Average Precision (mAP) of 94.6%. Similarly, the application of vertical flip (T03) results in a more noticeable drop of 1.2%, with an mAP of 93.9%. More substantial performance degradation occurs with grayscale conversion (T04), where the model's mAP sharply declines by 13.6% to 81.5%. In contrast, using auto contrast (T05) or Gaussian blur (T06) results in minimal changes, with mAP values of 94.8%, showing slight decreases of 0.3% and 0.2%, respectively.

Overall, these results highlight that current DA techniques lead to substantial performance losses, emphasizing the effectiveness of the proposed DA strategies for training the QACL-Net.

4.3.3. Ablation Experiments on Mosaic Techniques

Table 6 presents an analysis of the impact of varying EQM and classic Mosaic probabilities on the mAP performance of the QACL-Net-B3 model on the NWPU10 dataset. Specifically, we kept the classic Mosaic technique consistently active to analyze its impact on existing methods.

Table 6. Impact of DA techniques on mAP (%) Performance of QACL-Net-B3

Model	EQM Probability				Classic Mosaic Probability	NWPU10
	0.1	0.3	0.5	0.8	1.0	mAP(%)
QACL-Net-B3	✓	×	×	×	×	95.1
	×	✓	×	×	×	95.0↓0.1
	×	×	✓	×	×	94.6↓0.5
	×	×	×	✓	×	92.9↓2.2
	×	×	×	×	✓	84.7↓10.4

As shown in Table 6, the proposed QACL-Net-B3 achieves an mAP of 95.1% when the EQM probability is set to 0.1. The model's performance remains nearly unchanged with an mAP of 95.0%, indicating a minimal decline of 0.1% when the EQM probability is increased to 0.3. However, as the EQM probability continues to rise, a more significant performance drop is observed. For instance, setting the EQM probability to 0.5 results in an mAP decrease to 94.6%, reflecting a reduction of 0.5%. Further increases in the probability to 0.8 lead to a more noticeable drop to 92.9%, a decline of 2.2%. In contrast, when the classic Mosaic probability is set to 1.0, the model experiences a substantial decrease to 84.7%, a reduction of 10.4%. These results indicate that while a lower EQM probability has minimal impact on model performance, higher probabilities—particularly for the classic Mosaic technique—lead to significant degradation in mAP.

4.4. Evaluation of Inference Efficiency

Table 7 presents the inference speeds and parameters of the QACL-Net-B0 and QACL-Net-B3 detectors. The experiments were conducted on a single NVIDIA 4070 Ti GPU using an image resolution of 640 × 640 pixels. For a more intuitive comparison, the Faster R-CNN model with a ResNet50-FPN backbone was utilized as the baseline.

Table 7. Comparison of Inference Speeds and Parameters for QACL-Net

Method	Backbone	Parameters (M)	FLOPs (G)	Frames Per Second (FPS)
Faster R-CNN-R50-FPN (Baseline)	ResNet-50	41.8	91.2	73
QACL-Net-B0 (Ours)	EfficientNet-B0	21.2	58.2	116
QACL-Net-B3 (Ours)	EfficientNet-B3	27.9	63.3	85

Notably, QACL-Net-B3, which employs the EfficientNet-B3 backbone, demonstrates a parameter count of 27.9 million and a FLOPs value of 63.3 billion, resulting in an inference speed of 85 frames per second (FPS). In contrast, the Faster R-CNN baseline, with a significantly higher parameter count of 41.8 million and FLOPs of 91.2 billion, achieves an inference speed of 73 FPS, indicating that QACL-Net-B3 is not only more efficient in terms of parameters but also offers a superior FPS rate than the baseline. Moreover, QACL-Net-B0 outperforms QACL-Net-B3 in terms of inference speed, reaching 116 FPS with only 21.2 million parameters and 58.2 billion FLOPs. This suggests that while QACL-Net-B3 provides a balanced trade-off between accuracy and computational efficiency, QACL-Net-B0 excels in speed with a smaller model size. Overall, the performance of QACL-Net highlights its competitive advantage in resource efficiency relative to traditional architectures while maintaining a robust inference speed.

4.5. Visualization and Analysis

4.5.1. Object Detection Results Visualization

The detection results presented in Figure 7 reveal that QACL-Net-B3 demonstrates a robust capability to accurately detect various objects in diverse RSIs. Notably, the model excels in identifying small targets, such as vehicles, which are often challenging to detect due to their size and the complexity of the surrounding environment. However, the model still makes an incorrect detection for the dam's gates, as illustrated in the bottom right corner of Figure 7, where these structures are misclassified as ships due to their highly similar appearances. Despite this limitation, QACL-Net-B3 consistently showcases its ability to discern and localize objects with various shapes.

Figure 7. Visualization of Detection Results for QACL-Net-B3 Across Different Objects

4.5.2. Class Activation Mapping for Visual Interpretability

Figure 8 illustrates the class activation mappings (CAMs) for the same RSI samples depicted in Figure 7, utilizing the gradient-weighted class activation mapping (Grad-CAM) technique.

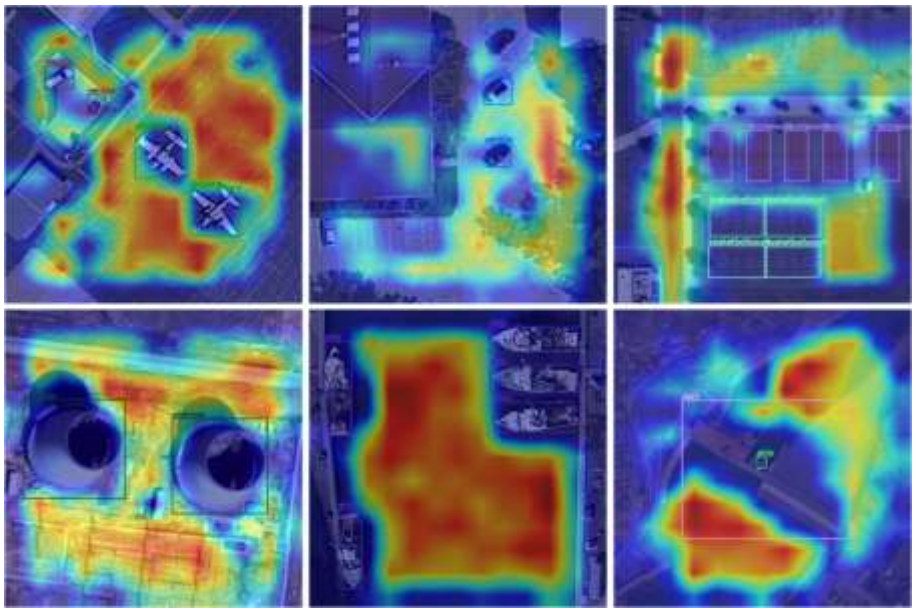

Figure 8. Visualizing Class Activation Mapping on Representative RSI Samples

The data shown in Figure 8 indicates that the CAMs produced by QACL-Net-B3 are significantly influenced by the features of the ground scenes. For example, the upper three subplots highlight extensive bright areas that correspond to important information about concrete surfaces, which are critical for the detector's predictions. Conversely, in scenes where objects are situated within water bodies—illustrated in the two lower right subplots—the model's activation areas are predominantly focused around the water. Additionally, when scenes contain multiple objects of similar sizes, such as the tennis court and chimneys featured in Figure 8, the bright activation areas become more scattered across the background. Overall, these findings suggest that the detector's CAMs are closely aligned with the semantic content of the ground scenes, emphasizing the benefits of multi-scale feature fusion in enhancing detection performance.

5. Conclusion

Recent advancements in object detection models have largely overlooked key characteristics of RSIs, leading to various suboptimal approaches that fail to strike an effective balance between model accuracy and computational efficiency. To the authors' knowledge, the inherent scarcity and complexity of RSI samples are central to these limitations, suggesting that modifying the model architecture alone cannot fully address the issue.

To address these challenges, we propose QACL-Net, an object detection framework built upon the Faster R-CNN architecture. This approach significantly enhances the performance of CNN-based detectors for RSI recognition while maintaining fast inference speeds. QACL-Net incorporates several novel techniques. First, we introduce the QA strategy to alleviate the issue of RSI sample scarcity. Second, we propose the EQM algorithm to improve the classic mosaic technique, which has proven less effective for RSI detection. Third, we introduce the CL strategy to resolve the problem of redundant feature fusion in the FPN, an issue that has often been overlooked in prior studies. Finally, we develop two variants of QACL-Net, each utilizing a different EfficientNet backbone—EfficientNet-B0 and EfficientNet-B3—for the detector architecture.

Extensive experiments on two widely used RSI datasets show that QACL-Net has outperformed 31 methods since 2022 on the DIOR20 dataset. Specifically, QACL-Net-B3 achieves a 6.9% improvement in accuracy on the challenging DIOR20 dataset, significantly surpassing other top-performing models. Additionally, QACL-Net reduces model size by 33% and increases inference speed by 17% compared to the baseline model. In summary, our work highlights the significant impact of RSI sample scarcity and feature fusion redundancy on object detection performance. As a promising solution, we propose QACL-Net, which effectively balances accuracy and computational efficiency. Theoretically, our approach can be seamlessly integrated with other detection models, as the QA, EQM, and CL modules require only minimal modifications to the FPN structure.

Despite these encouraging results, QACL-Net remains a preliminary solution and may not fully capture the diversity of real-world RSI scenarios. Future work could explore more effective strategies for feature fusion, extend evaluation to a wider range of RSI datasets, and investigate the integration of our method with other backbone architectures.

An important direction for future research involves extending our approach to one-stage detection frameworks, such as YOLO models, to investigate whether the proposed method can maintain its efficacy in real-time detection scenarios. This extension will be critical in advancing QACL-Net toward more practical and scalable solutions for RSI detection.

CRediT Author Contribution Statement

Huaxiang Song: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Resources, Data Curation, Writing—Original Draft, Writing—Review & Editing, Visualization, Supervision, Project Administration, and Funding Acquisition; Junping Xie: Methodology, Software, and Writing—Review & Editing; Yan Zhang: Conceptualization, Methodology, and Writing—Review & Editing; Yang Zhou, Wenhui Wang, YingYing Duan, and Xinyi Xie: Investigation, Resources, and Data Curation.

Acknowledgement

This work was funded by the Hunan Provincial Department of Education's Scientific Research Project (Project No. 24A0482) and the Research Foundation of Hunan University of Arts and Science (Geography Subject [2022] 351).

References

[1] Shouhang Du, Xiuyuan Zhang, Yichen Lei, Xin Huang, Wei Tu *et al.*, "Mapping urban functional zones with remote sensing and geospatial big data: a systematic review", *GIScience & Remote Sensing*, Print ISSN: 1548-1603, Vol. 61, No. 1, 31 December 2024, p. 2404900, Published by Taylor & Francis, DOI: 10.1080/15481603.2024.2404900, Available: https://www.tandfonline.com/doi/full/10.1080/15481603.2024.2404900.

[2] Gordana Kaplan, Fatma Yalcinkaya, Esra Altıok, Andrea Pietrelli, Rosa Anna Nastro *et al.*, "The role of remote sensing in the evolution of water pollution detection and monitoring: A comprehensive review", *Physics and Chemistry of the Earth, Parts A/B/C*, Print ISSN: 1474-7065, Vol. 136, 1 December 2024, p. 103712, DOI: 10.1016/j.pce.2024.103712, Available: https://www.sciencedirect.com/science/article/pii/S1474706524001700.

[3] Liegang Xia, Ruiyan Liu, Yishao Su, Shulin Mi, Dezhi Yang *et al.*, "Crop field extraction from high resolution remote sensing images based on semantic edges and spatial structure map", *Geocarto International*, Print ISSN: 1010-6049, Vol. 39, No. 1, 1 January 2024, p. 2302176, Published by Taylor & Francis, DOI: 10.1080/10106049.2024.2302176, Available: https://www.tandfonline.com/doi/full/10.1080/10106049.2024.2302176.

[4] Ao Chen, Zehua Lv, Junbo Zhang, Gangyi Yu and Rong Wan, "Review of the Accuracy of Satellite Remote Sensing Techniques in Identifying Coastal Aquaculture Facilities", *Fishes*, Print ISSN: 2410-3888, Vol. 9, No. 2, 2024, p. 52, Published by MDPI, DOI: 10.3390/fishes9020052, Available: https://www.mdpi.com/2410-3888/9/2/52.

[5] Carlos Lara-Alvarez, Juan J. Flores, Hector Rodriguez-Rangel and Rodrigo Lopez-Farias, "A literature review on satellite image time series forecasting: Methods and applications for remote sensing", *WIREs Data Mining and Knowledge Discovery*, Print ISSN: 1942-4787, Vol. 14, No. 3, 1 May 2024, p. e1528, Published by John Wiley & Sons, Ltd, DOI: 10.1002/widm.1528, Available: https://doi.org/10.1002/widm.1528.

[6] Wandong Jiang, Yuli Sun, Lin Lei, Gangyao Kuang and Kefeng Ji, "Change detection of multisource remote sensing images: a review", *International Journal of Digital Earth*, Print ISSN: 1753-8947, Vol. 17, No. 1, 31 December 2024, p. 2398051, Published by Taylor & Francis, DOI: 10.1080/17538947.2024.2398051, Available: https://www.tandfonline.com/doi/full/10.1080/17538947.2024.2398051.

[7] Yansheng Li, Junwei Luo, Yongjun Zhang, Yihua Tan, Jin-Gang Yu *et al.*, "Learning to Holistically Detect Bridges From Large-Size VHR Remote Sensing Imagery", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Print ISSN: 0162-8828, 2160-9292, 1939-3539, Vol. 46, No. 12, December 2024, pp. 11507–11523, Published by IEEE, DOI: 10.1109/TPAMI.2024.3393024, Available: https://ieeexplore.ieee.org/document/10509806/.

[8] Yixia Chen, Mingwei Lin, Zhu He, Kemal Polat, Adi Alhudhaif *et al.*, "Consistency- and dependence-guided knowledge distillation for object detection in remote sensing images", *Expert Systems with Applications*, Print ISSN: 0957-4174, Vol. 229, 1 November 2023, p. 120519, Published by Elsevier, DOI: 10.1016/j.eswa.2023.120519, Available: https://www.sciencedirect.com/science/article/pii/S0957417423010217.

[9] Chetan M Badgujar, Alwin Poulose and Hao Gan, "Agricultural object detection with You Only Look Once (YOLO) Algorithm: A bibliometric and systematic literature review", *Computers and Electronics in Agriculture*, Print ISSN: 0168-1699, Vol. 223, 1 August 2024, p. 109090, Published by Elsevier, DOI: 10.1016/j.compag.2024.109090, Available: https://www.sciencedirect.com/science/article/pii/S0168169924004812.

[10] Bowen Chen, Liqin Liu, Zhengxia Zou and Zhenwei Shi, "Target Detection in Hyperspectral Remote Sensing Image: Current Status and Challenges", *Remote Sensing*, Print ISSN: 2072-4292, Vol. 15, No. 13, 4 May 2023, p. 3223, Published by MDPI, DOI: 10.3390/rs15133223, Available: https://www.mdpi.com/2072-4292/15/13/3223.

[11] Jiaqi Wang, Zhihui Gong, Xiangyun Liu, Haitao Guo, Jun Lu *et al.*, "Multi-Feature Information Complementary Detector: A High-Precision Object Detection Model for Remote Sensing Images", *Remote Sensing*, Print ISSN: 2072-4292, Vol. 14, No. 18, 9 September 2022, p. 4519, Published by MDPI, DOI: 10.3390/rs14184519, Available: https://www.mdpi.com/2072-4292/14/18/4519.

[12] Nanjing Yu, Haohao Ren, Tianmin Deng and Xiaobiao Fan, "Stepwise Locating Bidirectional Pyramid Network for Object Detection in Remote Sensing Imagery", *IEEE Geoscience and Remote Sensing Letters*, Print ISSN: 1545-598X, Vol. 20, 2023, pp. 1–5, Published by IEEE, DOI: 10.1109/LGRS.2022.3223470, Available: https://ieeexplore.ieee.org/document/9955559/.

[13] Chengcheng Fan and Zhiruo Fang, "Probability-Enhanced Anchor-Free Detector for Remote-Sensing Object Detection", *Computers, Materials & Continua*, Print ISSN: 1546-2226, Vol. 79, No. 3, 2024, pp. 4925–4943, Published by Tech Science Press, DOI: 10.32604/cmc.2024.049710, Available: https://www.techscience.com/cmc/v79n3/57111.

[14] Shu Tian, Lin Cao, Lihong Kang, Xiangwei Xing, Jing Tian *et al.*, "A Novel Hybrid Attention-Driven Multistream Hierarchical Graph Embedding Network for Remote Sensing Object Detection", *Remote Sensing*, Print ISSN: 2072-4292, Vol. 14, No. 19, 4 October 2022, p. 4951, Published by MDPI, DOI: 10.3390/rs14194951, Available: https://www.mdpi.com/2072-4292/14/19/4951.

[15] Dongjun Zhu, Shixiong Xia, Jiaqi Zhao, Yong Zhou, Qiang Niu *et al.*, "Spatial hierarchy perception and hard samples metric learning for high-resolution remote sensing image object detection", *Applied Intelligence*, Print ISSN: 0924-669X, Vol. 52, No. 3, February 2022, pp. 3193–3208, Published by Springer, DOI: 10.1007/s10489-021-02335-0, Available: https://link.springer.com/10.1007/s10489-021-02335-0.

[16] Jiahang Liu, Donghao Yang and Fei Hu, "Multiscale Object Detection in Remote Sensing Images Combined with Multi-Receptive-Field Features and Relation-Connected Attention", *Remote Sensing*, Print ISSN: 2072-4292, Vol. 14, No. 2, 17 January 2022, p. 427, Published by MDPI, DOI: 10.3390/rs14020427, Available: https://www.mdpi.com/2072-4292/14/2/427.

[17] Huaxiang Song, Yuxuan Yuan, Zhiwei Ouyang, Yu Yang and Hui Xiang, "Quantitative regularization in robust vision transformer for remote sensing image classification", *The Photogrammetric Record*, Print ISSN: 0031-868X, Vol. 39, No. 186, June 2024, p. 340–372, Published by John Wiley & Sons Ltd, DOI: 10.1111/phor.12489, Available: https://onlinelibrary.wiley.com/doi/10.1111/phor.12489.

[18] Yansheng Li, Linlin Wang, Tingzhu Wang, Xue Yang, Junwei Luo *et al.*, "STAR: A First-Ever Dataset and a Large-Scale Benchmark for Scene Graph Generation in Large-Size Satellite Imagery", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Print ISSN: 0162-8828, 2160-9292, 1939-3539, Vol. 47, No. 3, March 2025, pp. 1832–1849, Published by IEEE, DOI: 10.1109/TPAMI.2024.3508072, Available: https://ieeexplore.ieee.org/document/10770756/.

[19] Huaxiang Song, "FST-EfficientNetV2: Exceptional Image Classification for Remote Sensing", *Computer Systems Science and Engineering*, Print ISSN: 0267-6192, Vol. 46, No. 3, 2023, pp. 3959–3978, Published by Tech Science Press, DOI: 10.32604/csse.2023.038429, Available: https://www.techscience.com/csse/v46n3/52217.

[20] Daifeng Peng, Xuelian Liu, Yongjun Zhang, Haiyan Guan, Yansheng Li *et al.*, "Deep learning change detection techniques for optical remote sensing imagery: Status, perspectives and challenges", *International Journal of Applied Earth Observation and Geoinformation*, Print ISSN: 1569-8432, Vol. 136, 1 February 2025, p. 104282, DOI: 10.1016/j.jag.2024.104282, Available: https://www.sciencedirect.com/science/article/pii/S1569843224006381.

[21] A.S.M. Sharifuzzaman Sagar, Yu Chen, YaKun Xie and Hyung Seok Kim, "MSA R-CNN: A comprehensive approach to remote sensing object detection and scene understanding", *Expert Systems with Applications*, Print ISSN: 0957-4174, Vol. 241, May 2024, p. 122788, Published by Elsevier, DOI: 10.1016/j.eswa.2023.122788, Available: https://linkinghub.elsevier.com/retrieve/pii/S0957417423032906.

[22] Zhe Guo, Guoling Bi, Hengyi Lv, Yuchen Zhao and Lintao Han, "Semantic Information Feature Aggregation Network for Object Detection in Remote Sensing Images", *IEEE Geoscience and Remote Sensing Letters*, Print ISSN: 1545-598X, Vol. 21, 2024, pp. 1–5, Published by IEEE, DOI: 10.1109/LGRS.2024.3406345, Available: https://ieeexplore.ieee.org/document/10540109/.

[23] Sagar A. S. M. Sharifuzzaman, Jawad Tanveer, Yu Chen, Jun Hoong Chan, Hyung Seok Kim *et al.*, "Bayes R-CNN: An Uncertainty-Aware Bayesian Approach to Object Detection in Remote Sensing Imagery for Enhanced Scene Interpretation", *Remote Sensing*, Print ISSN: 2072-4292, Vol. 16, No. 13, 30 June 2024, p. 2405, Published by MDPI, DOI: 10.3390/rs16132405, Available: https://www.mdpi.com/2072-4292/16/13/2405.

[24] Yanfeng Liu, Qiang Li, Yuan Yuan, Qian Du and Qi Wang, "ABNet: Adaptive Balanced Network for Multiscale Object Detection in Remote Sensing Imagery", *IEEE Transactions on Geoscience and Remote Sensing*, Print ISSN: 0196-2892, Vol. 60, 2022, pp. 1–14, Published by IEEE, DOI: 10.1109/TGRS.2021.3133956, Available: https://ieeexplore.ieee.org/document/9643004/.

[25] Guangming Xu, Tiecheng Song, Xia Sun and Chenqiang Gao, "TransMIN: Transformer-Guided Multi-Interaction Network for Remote Sensing Object Detection", *IEEE Geoscience and Remote Sensing Letters*, Print ISSN: 1545-598X, Vol. 20, 2023, pp. 1–5, Published by IEEE, DOI: 10.1109/LGRS.2022.3230973, Available: https://ieeexplore.ieee.org/document/9994788/.

[26] Huaxiang Song, Chai Wei and Zhou Yong, "Efficient knowledge distillation for remote sensing image classification: a CNN-based approach", *International Journal of Web Information Systems*, Print ISSN: 1744-0084, Vol. 20, No. 2, 14 December 2023, pp. 129–158, Published by Emerald Publishing Limited, DOI: 10.1108/IJWIS-10-2023-0192, Available: https://www.emerald.com/insight/content/doi/10.1108/IJWIS-10-2023-0192/full/html.

[27] Huabin Diao, Gongyan Li, Shaoyun Xu, Chao Kong, Wei Wang *et al.*, "Self-distillation enhanced adaptive pruning of convolutional neural networks", Pattern Recognition, Print ISSN: 0031-3203, Vol. 157, 1 January 2025, p. 110942, Published by Elsevier, DOI: 10.1016/j.patcog.2024.110942, Available: https://www.sciencedirect.com/science/article/pii/S0031320324006939.

[28] Peng Yang, Dashuai Yu and Guowei Yang, "Object detection in aerial remote sensing images using bidirectional enhancement FPN and attention module with data augmentation", *Multimedia Tools and Applications*, Print ISSN: 1573-7721, Vol. 83, No. 13, 5 October 2023, pp. 38635–38656, Published by Springer, DOI: 10.1007/s11042-023-16973-8, Available: https://link.springer.com/10.1007/s11042-023-16973-8.

[29] Yalun Dai, Fei Ma, Wei Hu and Fan Zhang, "SPGC: Shape-Prior-Based Generated Content Data Augmentation for Remote Sensing Object Detection", *IEEE Transactions on Geoscience and Remote Sensing*, Print ISSN: 0196-2892, Vol. 62, 2024, pp. 1–11, Published by IEEE, DOI: 10.1109/TGRS.2024.3373442, Available: https://ieeexplore.ieee.org/document/10459239/.

[30] Fangli Mou, Zide Fan, Chuan'ao Jiang, Yidan Zhang, Lei Wang *et al.*, "Double Augmentation: A Modal Transforming Method for Ship Detection in Remote Sensing Imagery", *Remote Sensing*, Print ISSN: 2072-4292, Vol. 16, No. 3, 5 February 2024, p. 600, Published by MDPI, DOI: 10.3390/rs16030600, Available: https://www.mdpi.com/2072-4292/16/3/600.

[31] Xiong Xu, Beibei Zhao, Xiaohua Tong, Huan Xie, Yongjiu Feng *et al.*, "A Data Augmentation Strategy Combining a Modified pix2pix Model and the Copy-Paste Operator for Solid Waste Detection With Remote Sensing Images", *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, Print ISSN: 1939-1404, Vol. 15, 2022, pp. 8484–8491, Published by IEEE, DOI: 10.1109/JSTARS.2022.3209967, Available: https://ieeexplore.ieee.org/document/9904838/.

[32] Weixing Liu, Bin Luo and Jun Liu, "Synthetic Data Augmentation Using Multiscale Attention CycleGAN for Aircraft Detection in Remote Sensing Images", *IEEE Geoscience and Remote Sensing Letters*, Print ISSN: 1545-598X, Vol. 19, 2022, pp. 1–5, Published by IEEE, DOI: 10.1109/LGRS.2021.3052017, Available: https://ieeexplore.ieee.org/document/9337932/.

[33] Xuesen Ma, Jindian Dong, Weixin Wei, Biao Zheng, Ji Ma *et al.*, "Remote Sensing Image Object Detection by Fusing Multi-Scale Contextual Features and Channel Enhancement", *In Proceedings of the 2023 International Joint Conference on Neural Networks (IJCNN)*, 18-23 June 2023, Gold Coast, Australia, ISBN: 978-1-66548-867-9, pp. 01–07, Published by IEEE, DOI: 10.1109/IJCNN54540.2023.10191739, Available: https://ieeexplore.ieee.org/document/10191739/.

[34] Zheng Li, Yongcheng Wang, Dongdong Xu, Yunxiao Gao and Tianqi Zhao, "TBNet: A texture and boundary-aware network for small weak object detection in remote-sensing imagery", *Pattern Recognition*, Print ISSN: 00313203, Vol. 158, February 2025, p. 110976, Published by Elsevier, DOI: 10.1016/j.patcog.2024.110976, Available: https://linkinghub.elsevier.com/retrieve/pii/S0031320324007271.

[35] Qingyun Li, Yushi Chen and Ying Zeng, "Transformer with Transfer CNN for Remote-Sensing-Image Object Detection", *Remote Sensing*, Print ISSN: 2072-4292, Vol. 14, No. 4, 17 February 2022, p. 984, Published by MDPI, DOI: 10.3390/rs14040984, Available: https://www.mdpi.com/2072-4292/14/4/984.

[36] Xiaoxu Feng, Xiwen Yao, Gong Cheng, Jungong Han and Junwei Han, "SAENet: Self-Supervised Adversarial and Equivariant Network for Weakly Supervised Object Detection in Remote Sensing Images", *IEEE Transactions on Geoscience and Remote Sensing*, Print ISSN: 0196-2892, Vol. 60, 2022, pp. 1–11, Published by IEEE, DOI: 10.1109/TGRS.2021.3105575, Available: https://ieeexplore.ieee.org/document/9524360/.

[37] Jinkang Wang, Xiaohui He, Shao Faming, Guanlin Lu, Qunyan Jiang *et al.*, "Multi-Size Object Detection in Large Scene Remote Sensing Images Under Dual Attention Mechanism", *IEEE Access*, Print ISSN: 2169-3536, Vol. 10, 2022, pp. 8021–8035, Published by IEEE, DOI: 10.1109/ACCESS.2022.3141059, Available: https://ieeexplore.ieee.org/document/9673732/.

[38] Zhu Teng, Yani Duan, Yan Liu, Baopeng Zhang and Jianping Fan, "Global to Local: Clip-LSTM-Based Object Detection From Remote Sensing Images", *IEEE Transactions on Geoscience and Remote Sensing*, Print ISSN: 0196-

2892, 1558-0644, Vol. 60, 2022, pp. 1–13, Published by IEEE, DOI: 10.1109/TGRS.2021.3064840, Available: https://ieeexplore.ieee.org/document/9386208/.

[39] Jiaojiao Li, Huanqing Zhang, Rui Song, Weiying Xie, Yunsong Li *et al.*, "Structure-Guided Feature Transform Hybrid Residual Network for Remote Sensing Object Detection", *IEEE Transactions on Geoscience and Remote Sensing*, Print ISSN: 0196-2892, 1558-0644, Vol. 60, 2022, pp. 1–13, Published by IEEE, DOI: 10.1109/TGRS.2021.3103964, Available: https://ieeexplore.ieee.org/document/9520130/.

[40] Pourya Shamsolmoali, Masoumeh Zareapoor, Jocelyn Chanussot, Huiyu Zhou and Jie Yang, "Rotation Equivariant Feature Image Pyramid Network for Object Detection in Optical Remote Sensing Imagery", *IEEE Transactions on Geoscience and Remote Sensing*, Print ISSN: 0196-2892, 1558-0644, Vol. 60, 2022, pp. 1–14, Published by IEEE, DOI: 10.1109/TGRS.2021.3112481, Available: https://ieeexplore.ieee.org/document/9547378/.

[41] Gong Cheng, Min He, Hailong Hong, Xiwen Yao, Xiaoliang Qian *et al.*, "Guiding Clean Features for Object Detection in Remote Sensing Images", *IEEE Geoscience and Remote Sensing Letters*, Print ISSN: 1545-598X, 1558-0571, Vol. 19, 2022, pp. 1–5, Published by IEEE, DOI: 10.1109/LGRS.2021.3104112, Available: https://ieeexplore.ieee.org/document/9515077/.

[42] Xiaohu Dong, Ruigang Fu, Yinghui Gao, Yao Qin, Yuanxin Ye *et al.*, "Remote Sensing Object Detection Based on Receptive Field Expansion Block", *IEEE Geoscience and Remote Sensing Letters*, Print ISSN: 1545-598X, 1558-0571, Vol. 19, 2022, pp. 1–5, Published by IEEE, DOI: 10.1109/LGRS.2021.3110584, Available: https://ieeexplore.ieee.org/document/9537586/.

[43] Tao Gao, Shilin Xia, Mengkun Liu, Jing Zhang, Ting Chen *et al.*, "MSNet: Multi-Scale Network for Object Detection in Remote Sensing Images", *Pattern Recognition*, Print ISSN: 00313203, Vol. 158, February 2025, p. 110983, Published by Elsevier, DOI: 10.1016/j.patcog.2024.110983, Available: https://linkinghub.elsevier.com/retrieve/pii/S0031320324007349.

[44] Kaihua Zhang and Haikuo Shen, "Multi-Stage Feature Enhancement Pyramid Network for Detecting Objects in Optical Remote Sensing Images", *Remote Sensing*, Print ISSN: 2072-4292, Vol. 14, No. 3, 26 January 2022, p. 579, Published by MDPI, DOI: 10.3390/rs14030579, Available: https://www.mdpi.com/2072-4292/14/3/579.

[45] Yuchao Zheng, Xinxin Zhang, Rui Zhang and Dahan Wang, "Gated Path Aggregation Feature Pyramid Network for Object Detection in Remote Sensing Images", *Remote Sensing*, Print ISSN: 2072-4292, Vol. 14, No. 18, 15 September 2022, p. 4614, Published by MDPI, DOI: 10.3390/rs14184614, Available: https://www.mdpi.com/2072-4292/14/18/4614.

[46] Luxuan Bian, Bo Li, Jue Wang and Zijun Gao, "Multi-branch stacking remote sensing image target detection based on YOLOv5", *The Egyptian Journal of Remote Sensing and Space Sciences*, Print ISSN: 11109823, Vol. 26, No. 4, December 2023, pp. 999–1008, Published by Elsevier, DOI: 10.1016/j.ejrs.2023.11.006, Available: https://linkinghub.elsevier.com/retrieve/pii/S1110982323000959.

[47] Yangguang Zhu, Xian Sun, Wenhui Diao, Hao Li and Kun Fu, "RFA-Net: Reconstructed Feature Alignment Network for Domain Adaptation Object Detection in Remote Sensing Imagery", *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, Print ISSN: 1939-1404, 2151-1535, Vol. 15, 2022, pp. 5689–5703, Published by IEEE, DOI: 10.1109/JSTARS.2022.3190699, Available: https://ieeexplore.ieee.org/document/9829266/.

[48] Huaxiang Song, "A More Efficient Approach for Remote Sensing Image Classification", *Computers, Materials & Continua*, Print ISSN: 1546-2226, Vol. 74, No. 3, 2023, pp. 5741–5756, Published by Tech Science Press, DOI: 10.32604/cmc.2023.034921, Available: https://www.techscience.com/cmc/v74n3/50961.