Journal of Multimedia Information System VOL. 10, NO. 3, September 2023 (pp. 215-226): ISSN 2383-7632 (Online)
https://doi.org/10.33851/JMIS.2023.10.3.215

A Method for Detecting Lightweight Optical Remote Sensing Images Using Improved Yolov5n

ChangMan Zou[1,2], Wang-Su Jeon[1], Sang-Yong Rhee[1*], MingXing Cai[1]

Abstract

Optical remote sensing image detection has wide-ranging applications in both military and civilian sectors. Addressing the specific challenge of false positives and missed detections in optical remote sensing image analysis due to object size variations, a lightweight remote sensing image detection method based on an improved YOLOv5n has been proposed. This technology allows for rapid and effective analysis of remote sensing images, real-time detection, and target localization, even in scenarios with limited computational resources in current machines/systems. To begin with, the YOLOv5n feature fusion network structure incorporates an adaptive spatial feature fusion mechanism to enhance the algorithm's ability to fuse features of objects at different scales. Additionally, an SIoU loss function has been developed based on the original YOLOv5n positional loss function, redefining the vector angle between position frame regressions and the penalty index. This adjustment aids in improving the convergence speed of model training and enhancing detection performance. To validate the effectiveness of the proposed method, experimental comparisons were conducted using optical remote sensing image datasets. The experimental results on optical remote sensing images serve to demonstrate the efficiency of this advanced technology. The findings indicate that the average mean accuracy of the improved network model has increased from the original 81.6% to 84.9%. Moreover, the average detection speed and network complexity are significantly superior to those of the other three existing object detection algorithms.

Key Words: Feature Fusion Mechanism, Object Detection, Remote Sensing Image, SIoU, Yolov5.

I. INTRODUCTION

Optical remote sensing image object detection is increasingly finding applications across various domains. It is widely utilized in civilian sectors, such as search and rescue operations, disaster monitoring and prediction, as well as urban construction planning. In the military domain, the detection and positioning of remote sensing objects enable the rapid conversion of remote sensing data into actionable intelligence. This capability proves invaluable in analyzing battlefield situations, accurately identifying the positions of potential targets, and subsequently formulating precise and timely military strategies [1]. As a result, achieving real-time and accurate detection holds significant importance, profoundly impacting both societal and economic development, as well as national defense efforts.

In recent years, deep learning techniques have gained significant traction among researchers tackling video generation and analysis tasks. These techniques involve using a preceding set of video frames to predict the subsequent set of frames within a given video sequence [2]. Some scholars have also leveraged image resolution enhancement in videos to facilitate local motion detection, allowing for the prompt identification of unwanted motion within the video content [3]. Inspired by advancements in video image detection algorithms, we aim to employ deep learning algorithms for object inspection and recognition in remote sensing images. This endeavor bears similarities to the individual frame detection commonly utilized in video analysis. Currently, mainstream remote sensing object detection algorithms predominantly fall into two categories [4-12]. In recent years, numerous scholars have dedicated their research efforts to this field. For instance, Xue Yali and Yao Qunli have proposed a lightweight object detection method tailored to enhancing the accuracy of identifying small objects amidst complex backgrounds in optical remote sensing images. This innovative approach tackles the challenges associated with detecting small objects, particularly when they are closely arranged. By incorporating a weighted fusion feature network, each layer's feature map receives a dynamically learned weight coefficient during network training, thus enhancing the fusion of deep and shallow

Manuscript received July 21, 2023; Revised August 28, 2023; Accepted September 11, 2023. (ID No. JMIS-23M-08-030)
Corresponding Author (*): Sang-Yong Rhee, +82-55-249-2706, syrhee@kyungnam.ac.kr
[1]Department of Computer Science and Engineering, Kyungnam University, Changwon, Korea, zouchangman@beihua.edu.cn, jws2218@naver.com, syrhee@kyungnam.ac.kr, caimingxinghh@naver.com
[2]College of Computer Science and Technology, Beihua University, Jilin, China, zouchangman@beihua.edu.cn

layer features. Moreover, the introduction of the CIoU loss function expedites network convergence, meeting real-time requirements [13]. Yao Qunli, in another study, has put forth a one-stage multi-scale feature fusion method designed for aircraft object detection, addressing the issue of low detection accuracy concerning small-scale aircraft objects in complex scenes. Regarding dataset utilization and processing, Vishal Pandey and colleagues have proposed several methods to enhance object detection in aerial images, promising substantial improvements in current aerial image detection performance [14].

While one-stage detection algorithms like YOLOv3, YOLOv4, and SSD offer faster detection speeds compared to two-stage detection algorithms, their network models tend to be relatively large and may not meet the practical lightweight deployment requirements. Previous research efforts have partly addressed the challenge of relatively low detection accuracy in one-stage algorithms by enhancing the network structure and employing various techniques. However, these enhancements often increased the network's complexity without achieving a satisfactory balance between detection accuracy and speed.

In light of these challenges, this paper introduces a lightweight multi-scale enhancement algorithm for remote sensing image detection. This approach effectively extracts and fuses features from remote sensing objects at different scales, addressing issues of errors and omissions in the detection process resulting from scale variations. Careful consideration is given to the trade-off between speed and accuracy in detection, resulting in a well-balanced approach.

To improve feature fusion at different scales, an adaptive spatial feature fusion mechanism is employed, leading to enhanced detection performance for remote sensing objects of varying sizes [15]. Additionally, the original algorithm's CIoU frame position loss function is replaced with the SIoU loss function [16]. The original CIoU loss function did not account for the mismatched direction between the required real frame and the predicted frame, which could lead to slow convergence and reduced detection efficiency. The SIoU loss function incorporates the vector angle between the real frame and the predicted frame, along with a redefined penalty index, thereby improving network training convergence speed and remote sensing image detection effectiveness.

Finally, the publicly available RSOD dataset [17] and NWPU VHR-10 [18] dataset was utilized as experimental data to evaluate the network's performance and compare it with other widely used object detection algorithms.

II. RELATED WORK

2.1. Feature Fusion

Feature fusion in object detection refers to the integration of features from various layers or modules within a network to enhance model performance and accuracy. The goal of feature fusion is to comprehensively leverage feature information at different levels to capture multi-scale object details, enrich contextual semantics, complement features across various levels or modules, and facilitate cross-layer feature propagation. There are several feature fusion methods in object detection, each with different variants and improvements across various research studies.

Among the commonly utilized feature fusion methods in object detection:Feature Pyramid Network (FPN) [19]: FPN is a widely adopted multi-scale feature fusion approach. It constructs a feature pyramid structure to facilitate cross-layer feature fusion. This method effectively captures semantic information from objects at multiple scales and provides rich contextual information. Pyramid Convolutional Neural Network (PANet) [20]: PANet is an enhanced version of the feature pyramid network that introduces both top-down and bottom-up feature propagation pathways. This modification better exploits contextual information between feature maps of varying scales. Deformable Convolutional Network (DCN) [21]: DCN is a feature fusion method designed to capture local object detail by learning adaptive deformable convolution kernels. It introduces spatial transformations in the feature extraction stage to adapt to object deformations and scale variations. Channel Attention Module (CAM) [22]: CAM is a feature fusion method that incorporates an attention mechanism. It adaptively adjusts the weight of each channel in the feature map to enhance the expression of essential features. Hybrid Feature Fusion Methods (e.g., BiFPN [23] and NAS-FPN [24]): These methods leverage diverse feature fusion strategies, including combining multi-scale feature pyramids with attention mechanisms. Such approaches significantly enhance the performance of object detection models. This article provides an overview of various feature fusion techniques in object detection, offering insights into their roles in improving model capabilities.

Adaptive Spatial Feature Fusion (ASFF) [25]: ASFF's primary concept revolves around dynamically adjusting feature weights based on the object's representation requirements across different spatial scales. This approach learns weight coefficients to adaptively combine multi-scale features, significantly improving the model's capacity to handle object detection at various scales. One notable advantage of this technique lies in its capability to dynamically fine-tune feature fusion weights, considering changes in object scale and contextual information. This adaptability enhances the detection of objects with varying scales. Following experimentation, we have opted for this adaptive feature fusion mechanism to bolster the performance and resilience of our object detection model, particularly when addressing tasks involving multi-scale objects. It is parti-

cularly well-suited for fulfilling the feature fusion demands of remote sensing image object detection discussed in this paper.

2.2. Loss Function

In the preceding section, we discussed feature fusion methods in object detection. Now, our focus shifts to the modification of the loss function within the network model. The YOLO series of algorithms introduced a transformative approach to object detection by framing it as a regression problem. This involves simultaneously predicting both the bounding box and category information of objects in a single forward pass. The YOLO series comprises multiple versions, each incorporating distinct loss functions and refinements. YOLOv1 primarily relies on two loss functions: bounding box regression loss and classification loss [26]. YOLOv2 builds upon YOLOv1 by introducing additional loss functions, including confidence loss, bounding box coordinate loss, category loss, and object loss [27]. Furthermore, YOLOv2 introduces multi-scale training and prediction, leveraging a more intricate grid division to enhance detection performance for smaller objects. YOLOv3 further refines the loss functions from YOLOv2 and introduces loss functions tailored to feature maps of varying scales.

Loss functions in object detection models encompass various components, including confidence loss, bounding box coordinate loss, category loss, object loss, and segmentation loss for occlusion detection [28]. YOLOv4 [29] and YOLOv5 [30] build upon this foundation by incorporating components such as confidence loss, bounding box coordinate loss, category loss, Landmark loss, and Focal Loss, among others. The specific implementations of YOLOv4 and YOLOv5 may exhibit subtle differences in their loss functions, depending on specific implementation details and the libraries utilized.

Furthermore, YOLOv5 introduces notable enhancements in the design of its loss function. It includes metrics like IoU (Intersection over Union), which primarily considers the overlapping area between the detection frame and the object frame. Building upon IoU, GIoU (Generalized-IoU) addresses bounding box alignment issues [31]. DIoU (Distance-IoU), an extension of IoU and GIoU, incorporates distance information from the bounding box's center point to enhance detection accuracy. Additionally, CIoU (Complete-IoU), based on DIOU, considers the aspect ratio of the bounding box's scale information, among other factors.

However, these loss functions primarily aggregate bounding box regression metrics, taking into account factors such as the distance between the predicted box and the ground truth box, overlapping area, and aspect ratio. Notably, the regression loss in the aforementioned models does not address the problem of direction mismatch, potentially leading to slower model convergence. During training, predicted boxes may oscillate around ground truth boxes, resulting in suboptimal results.

To address this issue, the SIoU (Smoothed IoU) loss function takes into account the vector angle between required regressions and redefines penalty indicators. These indicators encompass four components: angle cost, distance cost, shape cost, and IoU cost. This comprehensive approach significantly improves both training speed and inference accuracy.

III. REMOTE SENSING IMAGE DETECTION MODEL

In order to tackle the challenges associated with ship object classification and detection in high-resolution optical remote sensing images, this paper presents a novel network model for object detection, denoted as ASFF-SIoU-YOLOv5n (Adaptively Spatial Feature Fusion with SIoU-enhanced YOLOv5n) [32]. The overall structure of the proposed network model is illustrated in Fig. 1. To begin with, the foundational network, YOLOv5n, is employed as the basis for this model. YOLOv5n represents the version of YOLOv5 (version 6.0) with the smallest feature map width and network depth. Building upon this foundation, ASFF is seamlessly integrated into the YOLOv5n network architecture. This incorporation enhances the network's capacity to effectively fuse features at varying scales. Furthermore, a significant upgrade is made to the original YOLOv5n loss function, replacing it with the more advanced SIoU loss function (Smoothed IoU). This enhancement plays a pivotal role in achieving a delicate balance between lightweight deployment, high-speed processing, and high-precision remote sensing object detection.

3.1. Adaptive Feature Fusion Mechanism

The YOLOv5n object detection network implements the PANet structure to enhance the merging of multi-scale feature maps. PANet introduces a bottom-up refinement structure, which builds upon the existing FPN framework. It departs from the original single-item fusion approach, adopting a two-way fusion method. This design aims to leverage both the high-level semantic information present in optical remote sensing images and the fine-grained details found at the lower levels, such as contours, edges, colors, and shapes.

To fully exploit these diverse sources of information, the network incorporates an adaptive feature fusion mechanism

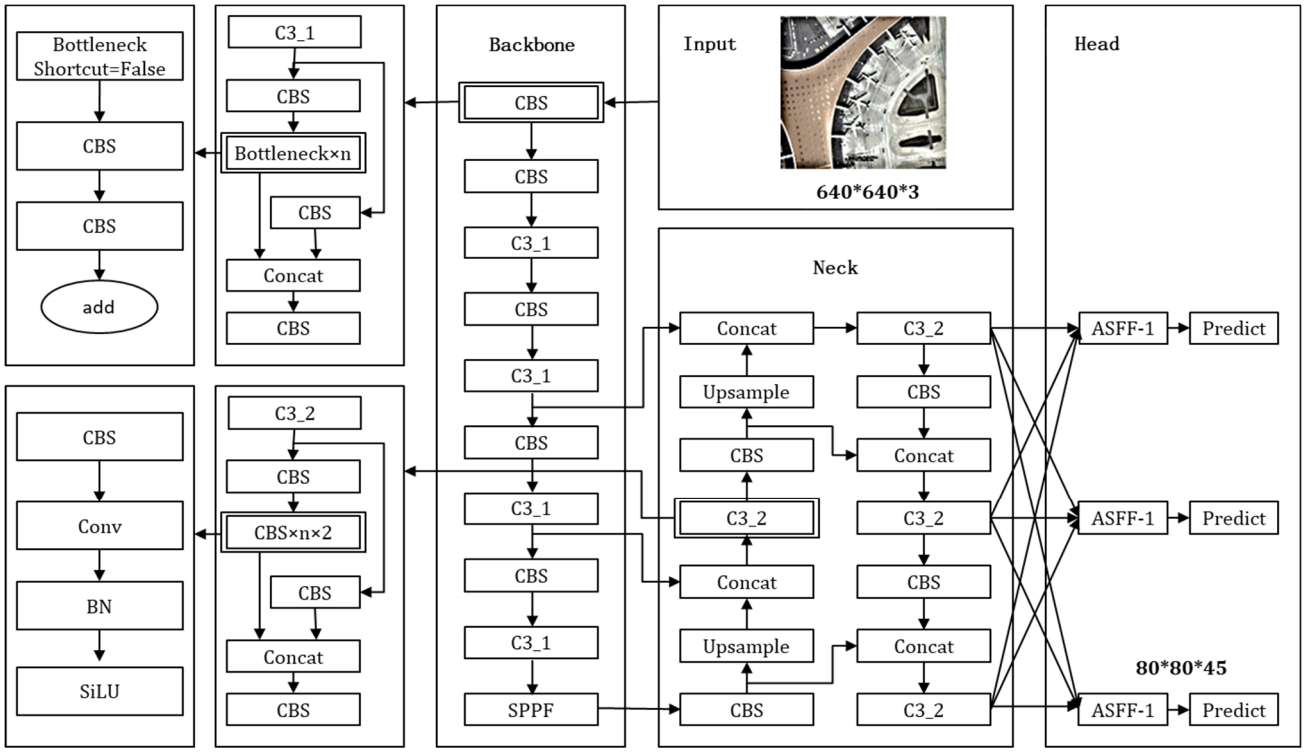


Fig. 1. AS-YOLOv5n network structure. The overall network structure of YOLOv5 consists of three parts: backbone is the main feature extraction network part, which mainly extracts image features, Neck is the feature fusion part, and Head is the detection part.

known as Adaptively Spatial Feature Fusion (ASFF). At the core of ASFF is the dynamic adjustment of weights during feature fusion across different scales. When combined with PANet, a fusion weight is learned for each layer scale. This adaptive weight allocation enables more effective utilization of features at different scales during the prediction of feature maps. Fig. 2 illustrates the structural framework of ASFF.

The feature fusion network output in YOLOv5n is the feature map of level1, level2 and level3. Taking ASFF-1 as an example, the fused output consists of semantic features from level 1, level 2, and level 3, along with the weight α obtained from different layers. β and γ are multiplied and added together. As shown in Equation (1):

$$y_{ij}^{1} = \alpha_{ij}^{1} \times x_{ij}^{1\to 1} + \beta_{ij}^{1} \times x_{ij}^{2\to 1} + \gamma_{ij}^{1} \times x_{ij}^{3\to 1}. \qquad (1)$$

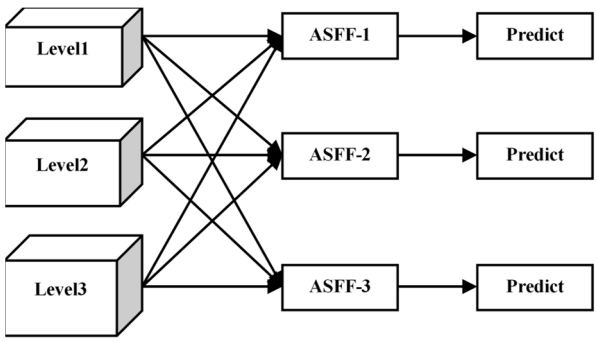


Fig. 2. ASFF structure diagram. The picture shows the cross-fusion between the three feature layers extracted from the backbone.

Among them, α_{ij}^{1}, β_{ij}^{1}, γ_{ij}^{1} are weights from different layers, $x_{ij}^{1\to 1}$, $x_{ij}^{2\to 1}$, $x_{ij}^{3\to 1}$ are outputs from different feature maps. Since the scale of the object to be measured in the remote sensing image varies widely, by introducing the ASFF method to learn the fusion method of the parameters, other less useful hierarchical features can be filtered, and only the useful information of this layer can be retained, thereby improving the accuracy of object detection.

3.2. Bounding Box Regression Loss Function Optimization

In computer vision tasks, the accuracy of object detection holds paramount importance, and this accuracy is significantly influenced by the choice of the loss function. In the original YOLOv5n detection algorithm, various metrics such as GIoU, CIoU, overlapping area, and aspect ratio are employed to calculate the loss function, primarily based on bounding box regression. However, a notable limitation of this approach is its failure to account for the direction mismatch between the predicted box and the ground truth box. This shortcoming leads to slower convergence and reduced efficiency in the training process.

To tackle this critical issue, Zhora introduces a novel loss function known as SIoU (Smoothed IoU). SIoU redefines the penalty metric by taking into consideration the vector angle between the required regressions. In the context of this paper, the original CIoU loss function is replaced with

SIoU to enhance the efficiency of object detection.

The SIoU loss function comprises four cost functions: angle cost, distance cost, shape cost, and IoU cost.

3.2.1. Angle Cost

The purpose of incorporating the angle-aware loss function component with the angle loss is to reduce the uncertainty associated with distance-related variables. Essentially, the model will prioritize aligning the prediction with either the X or Y axis (whichever is closer) before minimizing the distance along the corresponding axis.

Angle cost calculation formula is as follows:

$$\Lambda = 1 - 2 \times \sin^2(\arcsin(x) - \frac{\pi}{4}). \quad (2)$$

$$x = \frac{c_h}{\sigma} = sin(\alpha). \quad (3)$$

$$\sigma = \sqrt{(b_{c_x}^{gt} - b_{c_x})^2 + (b_{c_y}^{gt} - b_{c_y})^2}. \quad (4)$$

$$c_h = max\left(b_{c_y}^{gt} - b_{c_y}\right) - min(b_{c_y}^{gt} - b_{c_y}). \quad (5)$$

3.2.2. Distance Cost

The distance cost calculation formula is as follows:

$$\Delta = \sum_{t=x,y}(1 - e^{-\gamma\rho t}). \quad (6)$$

in addition,

$$\rho_x = (\frac{b_{c_x}^{gt} - b_{c_x}}{c_w})^2. \quad (7)$$

$$\rho_x = (\frac{b_{c_y}^{gt} - b_{c_y}}{c_h})^2. \quad (8)$$

$$\gamma = 2 - \Lambda. \quad (9)$$

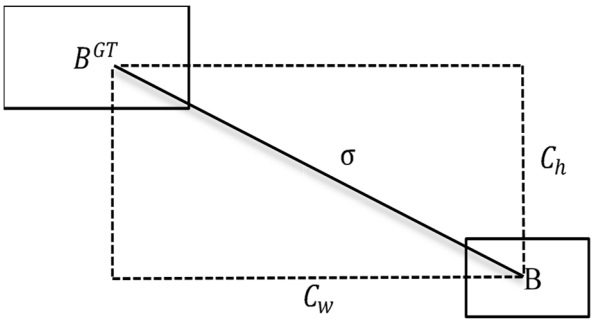


Fig. 3. The scheme for calculation of angle cost contribution into the loss function. Where $\mathbf{B^{GT}}$ and B are the center points of the prediction frame and the real frame. $\boldsymbol{\sigma}$ is the diagonal distance of the minimum circumscribed rectangle between the prediction frame and the real frame.

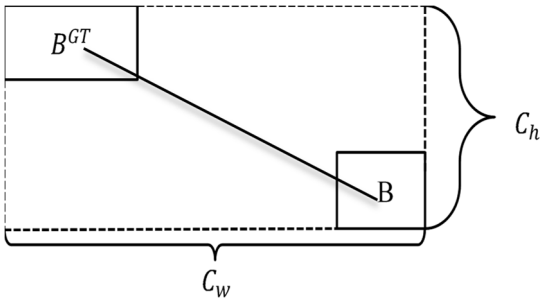


Fig. 4. Scheme for calculation of the distance between the ground truth bounding box and the prediction of it. Where $\boldsymbol{C_w}$ and $\boldsymbol{C_h}$ are the length and width of the outer rectangles of the real and prediction boxes, respectively.

3.2.3. The Shape Cost Calculation Formula is Defined as Follows

$$\Omega = \sum_{t=w,h}(1 - e^{-\omega t})^{\theta}. \quad (10)$$

$$\omega_w = \frac{|w - w^{gt}|}{max(w, w^{gt})}. \quad (11)$$

$$\omega_h = \frac{|h - h^{gt}|}{max(h, h^{gt})}. \quad (12)$$

3.2.4. The Regression Loss Function Expression of the Final Bounding Box is as Follows

$$L_{box} = 1 - IoU + \frac{\Omega + \Delta}{2}. \quad (13)$$

IV. EXPERIMENTAL DATA AND PROCESSING

The experimental data used for network model training in this paper comes from the RSOD dataset released by Wuhan University and NWPU VHR-10 dataset.

4.1. Evaluation Index

Before evaluating the model, it is very important to choose an appropriate evaluation metric.

The model's accuracy is evaluated using the recall rate (R), precision (P), average precision (AP), and average mean precision (mAP) metrics in this paper; the model weight and network parameters are used to evaluate the complexity of the network model. The network model becomes more complex as the value of the two increases. The specific calculation method of each indicator is as follows:

$$R = \frac{TP}{TP + FN}. \quad (14)$$

$$P = \frac{TP}{TP + FP}. \quad (15)$$

$$AP = \int_0^1 P(R)dR. \quad (16)$$

$$mAP = \frac{\sum_m AP}{m}. \quad (17)$$

4.2. Experiment Platform

This experiment is based on the Ubuntu 18.04 operating system, Intel (R) Xeon (R) Gold 5218 processor, 39G memory, 11 cores, using the Pytorch 1.8.0 framework, and a GeForce RTX 2080 Ti graphics card for network model training with 11GB of memory.

The Python version is 3.8 and the CUDA version is 11.1.1. The model training is set to 300 iterations with a batch size of 16. The learning rate is dynamically adjusted during the training process, and the NAG optimizer with a momentum of 0.937 is used for optimization. In the model training, the periodic learning rate is adjusted.

4.3. Dataset

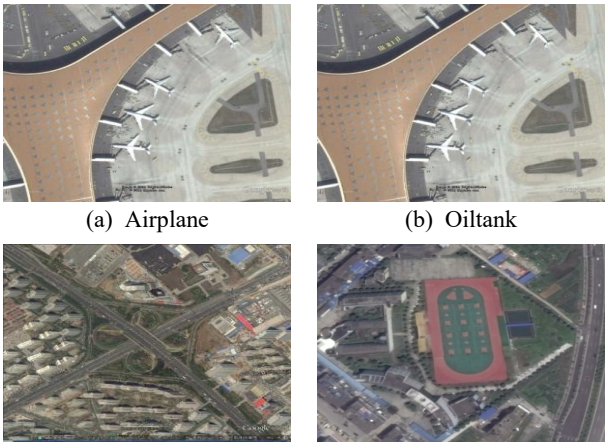
The experimental data used in the training of the network model in this study comes from the RSOD dataset [33] and NWPU VHR-10 [34] publicly available in China.

4.3.1. RSOD Dataset

The dataset used in this experiment comes from the domestic public RSOD dataset. The RSOD dataset has a total of 2326 images, and the dataset images are from Google Maps. The remote sensing dataset contains four categories of aircraft, oiltank, playgrounds, and overpasses. Among them, there are 446 images of aircraft, including 4,993 samples of aircraft; 165 images of oiltank, including 1,586 samples of oiltank; 189 images of playgrounds, including samples of playgrounds 191; 176 overpass images, including 180 overpass samples; the rest are background images. The dataset is divided randomly into training, validation, and test sets in a ratio of 7:1:2 in this paper. Fig. 5 shows some example images from this dataset. Fig. 5 visualizes the training progress of image classification and detection on the dataset. Fig. 6 shows the visualization of the image classification and detection training situation of the dataset.

4.3.2. NWPU VHR-10 Dataset

The second dataset used in this experiment comes from the public NWPU VHR-10 dataset. The dataset contains a total of 650 object images of 10 categories. The number of marked instances are 757 aircraft, 302 ships, 655 oil tanks, 390 baseball fields, 524 tennis courts, 159 basketball courts,163 track, field fields, 224 ports, 124 bridges and 477 vehicles. Fig. 7 shows some example images of the dataset.



(a) Airplane (b) Oiltank

(c) Overpass (d) Playgrond

Fig. 5. Visualization of partial RSOD datasets. Four images illustrate the types of remote sensing targets in the dataset.

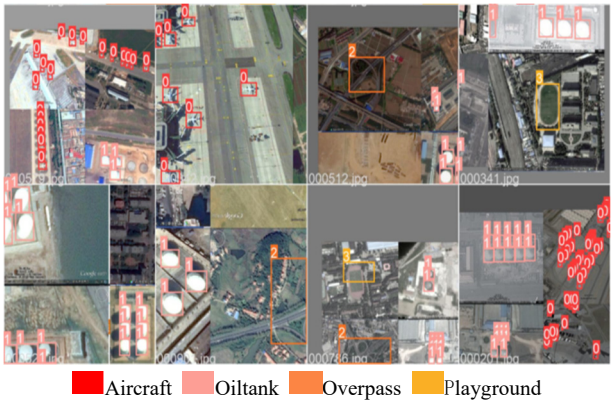
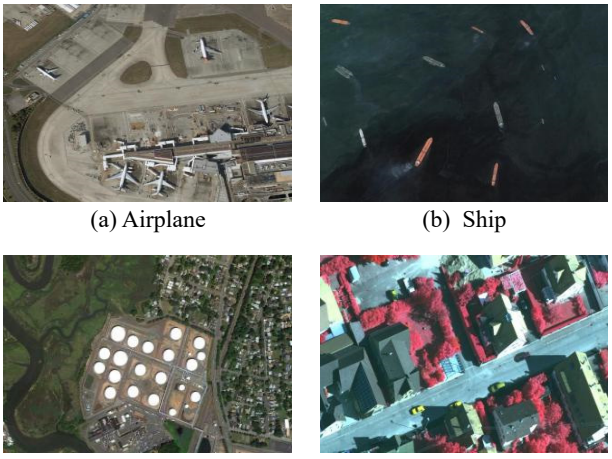


Fig. 6. Dataset image classification detection training visualization.



(a) Airplane (b) Ship

(c) Storage tank (d) Vehicle

Fig. 7. Visualization of partial NWPU VHR-10 datasets. Four images illustrate four types of remote sensing targets in the dataset.

V. EXPERIMENTAL RESULTS AND ANALYSIS

To assess the effectiveness of the AS-YOLOv5n algorithm introduced in this study, a series of experiments were conducted. These experiments involved comparing the AS-YOLOv5n with three other commonly used lightweight object detection algorithms. Additionally, the study sought to

investigate the individual contributions of each module within the algorithms discussed in this paper. To achieve this, ablation experiments were performed, particularly focusing on the improved ASFF and SIoU loss function.

All of the aforementioned experiments were carried out using the RSOD dataset for training the network model. Throughout the experiments, various factors such as equipment control, training hyperparameters, and the number of iterations were kept as fixed parameters. Subsequently, the acquired experimental results were thoroughly analyzed.

5.1. Compared with Other Methods

To validate the effectiveness of AS-YOLOv5n, comparative experiments were conducted using three target detection networks: YOLOv5n, YOLOv5s, and YOLOv3-Tiny. These experiments were performed on both the RSOD dataset and the NWPU VHR-10 dataset. During the training process, efforts were made to maintain as much consistency as possible in the parameters across the four network models. The training comprised 300 rounds, with an initial learning rate set at 0.01. The resulting experimental outcomes are presented in Table 1.

Table 1 highlights the remarkable performance of the proposed AS-YOLOv5n detection method, achieving an impressive mAP of 84.9% and 86.7% on the two datasets, respectively. Notably, on the RSOD dataset, AS-YOLOv5n outperforms other methods, with YOLOv3 yielding the highest mAP among the alternatives. The Tiny method shows a 1% improvement over the lowest YOLOv5n method, and it exceeds the lowest YOLOv5n method by 3.3%.

On the NWPU VHR-10 dataset, AS-YOLOv5n also excels, surpassing YOLOv5s, which yields the highest mAP among other methods, by 0.1%. Furthermore, AS-YOLOv5n outperforms the lowest-performing YOLOv3-Tiny method by a significant margin, with a 5.7% higher mAP.

Table 2. Comparison of experimental results on detection speed and network complexity on the RSOD dataset.

Methods	t/ms	Weight /MB	Gflops	Parameters
YOLOv5n	5.2	3.75	4.2	1764577
YOLOv5s	6.3	13.8	15.8	7020913
YOLOv3-Tiny	4.7	16.6	12.9	8673622
AS-YOLOv5n	5.4	6.42	6.3	3135594

Moreover, AS-YOLOv5n exhibits favorable AP values for each remote sensing object category, as evident from Table 2. It's worth noting that AS-YOLOv5n achieves these impressive results with considerably lower model parameters, weights, and computational resources compared to YOLOv5s and YOLOv3-Tiny. Furthermore, the time required to detect a single image is only 0.7 ms longer than YOLOv3-Tiny and 0.9 ms less than YOLOv5s.

Combining Tables 1 and Table 2, it can be seen that compared with other lightweight detection methods, the experimental results on the RSOD data set and NWPU VHR-10 data set show that the AS-YOLOv5n remote sensing image detection method proposed in this study achieves the highest mAP value, and the detection speed, model weight, model parameter amount and calculation amount are all excellent. Experimental comparison and verification show that the detection method proposed in this article balances detection speed and accuracy well. The AS-YOLOv5n algorithm surpasses other algorithms in terms of mAP. It also has the characteristics of simplicity and is suitable for actual deployment needs. The visualization of the detection effects of the four models is shown in Fig. 8.

5.2. Ablation Study

In order to verify the effectiveness of the introduction of the ASFF and SIoU methods proposed in this paper, three

Table 1. Experimental comparison results of average accuracy of different detection methods on the ground. The table lists the test results of two public data sets RSOD and NWPU VHR-10. Judging from the single type detection results, our model performed well. From the average accuracy it seems that our model is better than other models.

RSOD	AP1	AP2	AP3	AP4							mAP
YOLOv5n	97.8	98.2	83.3	47.1							81.6
YOLOv5s	**98.2**	**98.6**	82.3	44.8							83.5
YOLOv3-Tiny	96.3	96.3	76.8	**65.9**							83.9
AS-YOLOv5n	**98.2**	97.5	**86.8**	57.2							**84.9**
NWPU VHR-10	AP1	AP2	AP3	AP4	AP5	AP6	AP7	AP8	AP9	AP10	mAP
YOLOv5n	99.5	94.7	25.2	97.4	66.5	96.2	95.1	93.1	68.2	84.5	82.0
YOLOv5s	99.5	96.7	**70.9**	98.4	66.4	**97.2**	97.6	**93.2**	61.7	83.7	86.5
YOLOv3-Tiny	99.1	94.7	53.7	**98.9**	55.6	79.2	98.0	72.1	**82.9**	75.6	81.0
AS-YOLOv5n	**99.5**	**97.3**	61.4	97.8	**73.0**	88.4	**98.9**	90.5	71.3	**88.4**	**86.7**

The types in the ROSD data set include: AP1 (aircraft), AP2 (oiltank), AP3 (overpass), AP4 (playground).
The types in the NWPU VHR-10 data set include: AP1 (airplane), AP2 (ship), AP3 (storage tank), AP4(baseball diamond), AP5 (tennis court), AP6 (basketball court), AP7 (ground track field), AP8 (harbor), AP9 (bridge), AP10 (vehicle).

YOLOv5n YOLOv5s YOLOv3-Tiny AS-YOLOv5n

Fig. 8. Detection effect comparison chart，Different algorithms are used to check the same image, and the results show that our algorithm has certain advantages.

sets of experiments were compared. Table 3s and Table 4 present the results of ablation experiments conducted on the RSOD dataset, evaluating the performance of three different methods, Table 5 record the results of ablation experiments on the NWPU VHR-10 dataset, and obtain the comparison results of accuracy and network model complexity.

As shown in Table 3s and Table 5: After changing the loss function in the original YOLOv5n network model to SIoU, mAP increased the performance by 0.9% and 0.7% on the two data sets respectively. On this basis, we continued to introduce adaptive After the feature fusion method, the overall effect has been greatly improved, and the mAP value has increased by 2.4% and 4.0% respectively. As shown in Table 4, after improving the loss function, the model complexity did not change, but the detection speed was improved, and the single detection time was shortened by 0.2s. After the adaptive feature fusion method was introduced, the weight of the model increased by 2.67 MB. When the calculation amount and parameters are nearly doubled, the detection time of a single image only increases by 0.2 ms. It can be seen from Table 3 and Table 4 that after the SIoU loss function is introduced, the complexity of the model does not change, but the detection speed and detection accuracy are improved, indicating the effectiveness and advancement of the improved method. This shows that replacing the CIoU loss function of the original YOLOv5n with the SIoU position box regression loss function has been effectively verified; in addition, after the adaptive feature fusion method was introduced, the mAP obtained in the experiment and the AP value of each remote sensing target have greatly improved. The improvement, as shown in Table 5, also effectively verifies that the introduction of the adaptive feature fusion method has good performance in detecting objects of different scales in remote sensing

Table 3. Comparison of accuracy test experimental results on the RSOD data set.

Methods	mAP	AP_1	AP_2	AP_3	AP_4
YOLOv5n	81.6	97.8	**98.2**	83.3	47.1
YOLOv5n+SIoU	82.5	**98.5**	97.7	82.5	51.4
AS-YOLOv5n	**84.9**	98.2	97.5	**86.8**	**57.2**

There are four categories included in the data set: AP1 (aircraft), AP2 (oil tank), AP3 (overpass), AP4 (playground).

Table 4. Comparison of experimental results on detection speed and network complexity on the RSOD dataset,

Methods	t/ms	Weight/MB	Gflops	Parameters
YOLOv5n	5.2	3.75	4.2	1764577
YOLOv5n+SIoU	5.0	3.75	4.2	1764577
AS-YOLOv5n	**5.4**	**6.42**	**6.3**	**3135594**

VI. CONCLUSION

This paper presents a lightweight optical remote sensing image detection method based on an improved version of YOLOv5n. The method encompasses three key enhancements:

Firstly, it integrates ASFF into the YOLOv5n network structure, bolstering the network's capability to fuse features across different scales.

Secondly, the loss function of YOLOv5n is upgraded to the advanced SIoU, contributing to improved detection accuracy.

Finally, the proposed algorithm undergoes rigorous testing on a remote sensing image dataset and is benchmarked against three lightweight algorithms: YOLOv5n, YOLOv5s, and YOLOv3-Tiny. The experimental results clearly de-

Table 5. Comparison of experimental results on detection speed and network complexity on the NWPU VHR-10 dataset.

NWPU VHR-10	mAP	AP1	AP2	AP3	AP4	AP5	AP6	AP7	AP8	AP9	AP10
YOLOv5n	82.0	99.5	94.7	25.2	97.4	66.5	**96.2**	95.1	93.1	68.2	84.5
YOLOv5n+SIoU	82.7	99.5	95.8	16.6	97.2	67.5	91.5	98.7	90.3	**82.7**	87.3
AS-YOLOv5n	**86.7**	**99.5**	**97.3**	**61.4**	**97.8**	**73.0**	88.4	**98.9**	**90.5**	71.3	**88.4**

The types in the ROSD data set include: AP1 (aircraft), AP2 (oiltank), AP3 (overpass), AP4 (playground);
The types in the NWPU VHR-10 data set include: AP1 (airplane), AP2 (ship), AP3 (storage tank), AP4(baseball diamond), AP5 (tennis court), AP6 (basketball court), AP7 (ground track field), AP8 (harbor), AP9 (bridge), AP10 (vehicle).

monstrate that the enhanced network excels in accurately and efficiently detecting variations in remote sensing images.

Importantly, the proposed method significantly reduces errors and omissions compared to the original YOLOv5n algorithm. It outperforms traditional object detection algorithms in terms of detection speed, network model size, and accuracy. This method effectively addresses the challenges associated with remote sensing image detection, particularly erroneous and missed detections arising from scale variations.

Furthermore, this method fulfills the requirements for real-time and rapid detection of remote sensing objects, making it suitable for applications with limited computing resources and high-speed detection. It finds promising applications in scenarios like ocean search and rescue, maritime intelligence, reconnaissance, and early warning.

ACKNOWLEDGMENT

This work was supported by the Kyungnam region "SW Convergence Cluster 2.0 Specialized Industry" Reinforcement Project grant funded by the Ministry of Science and Technology Information and Communication (MSIT). This results was supported by "Regional Innovation Strategy (RIS)" through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (MOE) (2021 RIS-003).

This work is a phased research result of the Beihua University Education and Teaching Reform Research Project (No. XJQN20220019).

REFERENCES

[1] Y. L. Xue, Y. Sun, and H. R. Ma, "Lightweight small object detection in aerial remote sensing image", *Electronics Optics & Control*, vol. 29, no. 6, pp. 11-15, 2022.

[2] S. Mukherjee, S. Ghosh, S. Ghosh, P. Kumer, and P. P. Roy, "Predicting video-frames using encoder-convlstm combination," in *Proceedings of the ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 2019, pp. 2027-2031.

[3] S. Mukherjee, R. Saini, P. Kumar, P. P. Roy, D.P. Dogra, and B. G. Kim, "Fight detection in hockey videos using deep network", *Journal of Multimedia Information System*, vol. 4, no. 4, pp. 225-232, 2017.

[4] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580-587.

[5] R. Girshick, "Fast R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision(ICCV),* 2015, pp. 1440-1448.

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

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

[8] J. Redmon and A. Farhadi, "YOLOv9000: Better, faster, stronger," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition(CVPR),* Jul. 2017, pp. 6517-6525.

[9] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, and C. Y. Fu, et al., ″SSD: Single shot multibox detector,″ in *Proceedings of the 4th European Conferece on Computer Vision*, Switzerland, 2016, pp. 21-37.

[10] J. Redmon and A. Farhadi, ″YOLOv3: An incremental improvement,″ *arXiv Prepr. arXiv1804.02767*, 2018.

[11] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, *arXiv: 1804.0276*, 2018.

[12] Y. Liu, B. Lu, J. Peng, and Z. Zhang, "Research on the use of YOLOv5 object detection algorithm in mask wearing recognition," *World Scientific Research Journal*, vol. 6, no. 11, pp. 276-284, 2020.

[13] Y. A. O. Qunli, H. U. Xian, and L. Hong, "Aircraft detection in remote sensing imagery with multi-scale feature fusion convolutional neuralnetworks," *Acta Geodaetica et Cartographica Sinica*, vol. 48, no. 10, pp. 1266-1274, 2019.

[14] V. Pandey, K. Anand, A. Kalra, A. Gupta, P. P. Roy, and B. G. Kim, "Enhancing object detection in aerial images," *Mathematical Biosciences and Engineering*, vol. 19, no. 8, pp. 7920-7932, 2022.

[15] S. T. Liu, D. Huang, and Y. H. Wang, ″Learning spatial fusion for single-shot object detection,″ *arXiv Prepr. arXiv1911.09516*, 2019.

[16] Z. Gevorgyan, ″SIoU loss: More powerful learning for bounding box regression,″ *arXiv Prepr. arxiv: 2205. 12740*, 2022.

[17] GitHub, RSOD-Dataset, Apr. 2019, https://github.com/RSIA-LIESMARS-WHU/RSOD-Dataset-.

[18] Hyper, NWPU VHR-10 Geospatial Object Detection Remote Sensing Dataset, n.d. https://hyper.ai/datasets/

5422.

[19] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Feature pyramid networks for object detectionin," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2117-2125.

[20] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 8759-8768.

[21] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, and H. Hu, et al., "Deformable convolutional networksin," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV),* 2017, pp. 764-773.

[22] S. H. Woo, J. C. Park, J. Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV),* 2018, pp. 3-19.

[23] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* 2020, pp. 10781-10790.

[24] G. Ghiasi, T. Y. Lin, and Q. V. Le, "NAS-FPN: Learning scalable feature pyramid architecture for object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* 2019, pp. 7036-7045.

[25] S. Liu, D. Huang, and Y. Wang, "Learning spatial fusion for single-shot object detection*," arXiv Prep. arXiv:1911.09516* , 2019.

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

[27] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 7263-7271

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

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

[30] GitHub, ultralytics/yolov5, 2023, https://github.com/ultralytics/yolov5.

[31] H. Rezatofighi, N. Tsoi, J. Y. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* 2019, pp. 658-666.

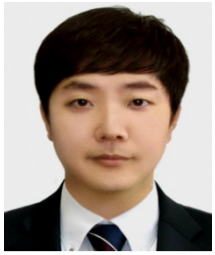
[32] S. Liu, D. Huang, and Y. H. Wang, "Learning spatial fusion for single-shot object detection," *arXiv Prepr. arXiv:1911.09516*, 2019.

[33] Y. Long, Y. P. Gong, Z. F. Xiao, and Q. Liu, "Accurate object localization in remote sensing images based on convolutional neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 5, pp. 2486-2498, 2017.

[34] G. Cheng, J. Han, and P, Zhou, and L. Guo, "Multiclassgeospatial object detection and geographic image classification based on collection of part detectors," *ISPRS Journal of Photogrammetry and Remote Sensing,* vol. 98, pp. 119-132, 2014.

AUTHORS

ChangMan Zou received his B.S. and M.S. degrees in computer science and technology from Beihua University, China, in 2008 and 2013, respectively. He is currently pursuing a Ph.D. at Kyungnam University, South Korea, since 2019. He is also a faculty member at the School of Computer Science and Technology at Beihua University. His research interests include deep learning and artificial intelligence.



Wang-Su Jeon received his B.S. and M.S. degrees in Computer Engineering and IT Convergence Engineering from Kyungnam University, Masan, Korea, in 2016 and 2018, and is currently pursuing the Ph.D. degree in IT Convergence Engineering at Kyungnam University, Masan, Korea. His present interests include computer vision.

Sang-Yong Rhee received his B.S. and M.S. degrees in Industrial Engineering from Korea University, Seoul, Korea, in 1982 and 1984, respectively, and his Ph.D. degree in Industrial Engineering at Pohang University, Pohang, Korea. He is currently a professor at the Computer Engineering, Kyungnam University, Masan, puter vision, augmented reality, human-robot interface.

MingXing Cai received his B.S. degree in software engineering from Kyungnam University, Korea, in 2023 and is currently pursuing a M.S. degree in the same university.