ISSN 1561-2430 (Print)
ISSN 2524-2415 (Online)

ИНФОРМАТИКА

INFORMATICS

UDC 004.93
https://doi.org/10.29235/1561-2430-2025-61-3-253-264

Received 07.08.2025
Поступила в редакцию 07.08.2025

Wu Xianyi[1], Sergey V. Ablameyko[1,2]

[1]Belarusian State University, Minsk, Republic of Belarus
[2]United Institute of Informatics Problems of the National Academy of Sciences of Belarus, Minsk, Belarus

A REMOTE SENSING IMAGE OBJECT DETECTION METHOD ABS-YOLO BASED ON IMPROVED YOLOv11

Abstract. The problem of object detection in Earth remote sensing images is studied, which is important for agricultural monitoring, urban planning, early warning of natural disasters, etc. Due to the different sizes of objects, complex background, and dense distribution of small objects in remote sensing images, problems such as high percentage of missed objects and insufficient accuracy of their coordinates often arise. In this regard, an improved method for YOLOv11, ABS-YOLO, is proposed, which significantly improves the performance of object detection by integrating Averaged Convolution (AConv), Bidirectional Weighted Feature Pyramid (BiFPN), and Swin Transformer attention mechanism. Experimental results show that, compared with YOLOv11, the proposed object detection method ABS-YOLO with AConv, BiFPN, and Swin Transformer achieves 3.9 % increase in mAP50 estimations and 2.6 % increase in mAP50-95 on the NWPU VHR-10 dataset with significant improvement in precision and recall rates. This method allows achieving a balance between efficiency and accuracy of remote sensing object detection due to the proposed improvements.

Keywords: YOLOv11, Swin Transformer, remote sensing image, object detection

For citation. Wu Xianyi, Ablameyko S. V. A remote sensing image object detection method ABS-YOLO based on improved YOLOv11. *Vestsi Natsyyanal'nai akademii navuk Belarusi. Seryya fizika-matematychnykh navuk = Proceedings of the National Academy of Sciences of Belarus. Physics and Mathematics series*, 2025, vol. 61, no. 3, pp. 253–264. https://doi.org/10.29235/1561-2430-2025-61-3-253-264

Ву Сяньи[1], С. В. Абламейко[1,2]

[1]Белорусский государственный университет, Минск, Республика Беларусь
[2]Объединенный институт проблем информатики Национальной академии наук Беларуси, Минск, Республика Беларусь

МЕТОД ОБНАРУЖЕНИЯ ОБЪЕКТОВ НА ИЗОБРАЖЕНИЯХ ДИСТАНЦИОННОГО ЗОНДИРОВАНИЯ ABS-YOLO НА ОСНОВЕ УЛУЧШЕННОГО YOLOv11

Аннотация. Исследуется задача обнаружения объектов на изображениях дистанционного зондирования Земли, что важно для сельскохозяйственного мониторинга, городского планирования, раннего предупреждения о стихийных бедствиях и др. Из-за различных размеров объектов, сложного фона и плотного распределения мелких объектов на изображениях дистанционного зондирования часто возникают проблемы, связанные с высоким процентом пропущенных объектов и недостаточной точностью определения их координат. В связи с этим предлагается усовершенствованный метод для YOLOv11 – ABS-YOLO, который значительно повышает производительность обнаружения объектов за счет интеграции усредненной свертки (AConv), двунаправленной пирамиды взвешенных признаков (BiFPN) и механизма внимания Swin Transformer. Экспериментальные результаты показывают, что по сравнению с YOLOv11 предложенный метод обнаружения объектов ABS-YOLO с AConv, BiFPN и Swin Transformer достигает увеличения оценок mAP50 на 3,9 % и mAP50-95 на 2,6 % на наборе данных NWPU VHR-10 со значительным улучшением в точности и показателях полноты. Данный метод позволяет достичь баланса между эффективностью и точностью обнаружения объектов дистанционного зондирования благодаря предложенным усовершенствованиям.

Ключевые слова: YOLOv11, Swin Transformer, изображение дистанционного зондирования, обнаружение объектов

Для цитирования. By Сяньи. Метод обнаружения объектов на изображениях дистанционного зондирования Земли ABS-YOLO на основе улучшенного YOLOv11 / By Сяньи, С. В. Абламейко// Весці Нацыянальнай акадэміі навук Беларусі. Серыя фізіка-матэматычных навук. – 2025. – Т. 61, № 3. – С. 253–264. https://doi.org/10.29235/1561-2430-2025-61-3-253-264

Introduction. Remote sensing object detection is one of the most fundamental and challenging tasks in the field of remote sensing, which has attracted people's attention for along time. Especially in recent years, with the development of deep learning technology, remote sensing object detection technology has made great progress. Advances in satellite technology have made the acquisition of remote sensing images (such as Google Earth [1]) and large-scale Earth observation [2] no longer difficult, even for high-resolution remote sensing images under changes in spatial, temporal, and spectral resolution [3]. However, remote sensing data detection still faces multiple challenges in the context of rapid technological development.

Traditional object detection methods mainly rely on manually designed feature extraction techniques, such as Histogram of Oriented Gradients (HOG) [4], Scale-Invariant Feature Transform (SIFT) [5], and Local Binary Pattern Histogram (LBPH) [6]. These methods show certain advantages in local feature extraction and classifier design, especially in scenes with simple structures and uniform backgrounds, where they are more applicable. However, when facing complex scenes, the discriminative ability of traditional methods is often limited, especially under diverse target shapes, textures, or lighting conditions, where their detection accuracy and robustness fail to meet practical requirements. In recent years, with the significant improvement of computer computing power, deep learning has gradually become an important research direction in the field of artificial intelligence. Among deep learning algorithms, object detection methods based on Convolutional Neural Networks (CNN) have achieved faster computational speed and higher detection accuracy due to their weight sharing and translational invariance, gradually replacing traditional methods and becoming the mainstream object detection technology today.

Object detection algorithms based on deep learning are mainly divided into two categories: two-stage detection and one-stage detection. Two-stage detection methods firstly extract features from the image and generate candidate regions, then classify and localize these candidate regions to ultimately output the position and category information of the target. Typical two-stage algorithms include R-CNN [7], Fast R-CNN [8], and Faster R-CNN [9]. These methods improve detection accuracy through staged processing but have higher computational complexity and are suitable for scenarios with high accuracy requirements. One-stage detection methods directly use CNN to extract image features and complete object detection through multiple fully connected layers without the need to pre-generate candidate regions. Their typical representatives include YOLO (You Only Look Once) [10] series and SSD (Single-Shot Multibox Detector) [11]. One-stage methods significantly improve detection efficiency by simplifying the process and eliminating the generation of candidate regions and frequent data transformations. They are particularly suitable for large-scale data processing tasks. Therefore, in practical applications, one-stage detection algorithms demonstrate broader application potential due to their high efficiency and real-time performance.

Although one-stage object detection algorithms represented by YOLO series have shown significant performance advantages in general scenes, they still face significant challenges in the field of remote sensing image analysis. The unique spatial characteristics of remote sensing images, including complex background interference, high-density target clustering, and a prominent proportion of sub-pixel small targets, lead to limitations in feature representation and localization accuracy for traditional one-stage detection paradigms. To enhance the detection performance of targets in remote sensing images, in recent years, many researchers have carried out multi-dimensional algorithm innovations around YOLO architecture. Yin Zhang et al. [12] proposed FFCA-YOLO, which enhances the weak feature representation of small targets and suppresses confusing backgrounds by integrating a Feature Enhancement Module (FEM), Feature Fusion Module (FFM), and Spatial Context-Aware Module (SCAM) into YOLO. Yi Hao et al. [13] proposed an improved YOLOv8 algorithm called LAR-YOLOv8, which uses a dual-branch architecture attention mechanism to enhance C2f module, reducing the repeated use of C2f module and achieving efficient feature extraction. Tianyong Wu et al. [14] proposed a new YOLOv8-based network called YOLO-SE, which integrates an Efficient Multi-scale Attention (EMA) mechanism

into the network to form an SPPFE module, addressing the issue of multi-scale target detection. Qiu Xue Wang et al. [15] proposed introducing SimSPPF module and dynamic large convolutional kernel attention mechanism LSK-attention into YOLOv8, optimizing the feature pyramid layer and expanding the model's receptive field to improve the accuracy of object detection. Shahriar Soudeep et al. [16] proposed an interpretable Dynamic Graph Neural Network: DGNN-YOLO, for small occluded object detection and tracking. The integration of Dynamic Graph Neural Network (DGNN) in YOLOv11 has improved detection accuracy and reliability.

From the work of the above researchers it can be seen that YOLO series algorithms have been widely applied to various object detection tasks, and there is still much room for improvement in specific domain detection tasks. This paper proposes a remote sensing image object detection algorithm based on improved YOLOv11n: ABS-YOLO (AConv-BiFPN-Swin Transformer-YOLOv11n). AConv (average pooling and convolution layer) is introduced in the shallow layer of the Backbone of this model. The average pooling operation is used to downsample the input feature map, reducing its size by half while retaining important feature information. Bidirectional Weighted Feature Pyramid (BiFPN) is introduced to improve the detection accuracy of small targets and complex scenes. Finally, Swin Transformer attention mechanism is integrated to suppress interference from complex backgrounds. On the basis of retaining the detection characteristics of YOLOv11n, ABS-YOLO model effectively improves accuracy and provides an efficient solution for remote sensing image object detection.

1. YOLOv11 Model. YOLOv11 [17], introduced by Ultralytics in 2024, represents a new generation of object detection algorithms with breakthrough innovations in architectural design. The backbone network utilizes an enhanced CSPDarknet architecture, optimizing cross-stage feature fusion via C3K2 module. This module employs dual small convolutional kernels instead of a single large kernel, significantly reducing computational redundancy. For feature fusion, YOLOv11 incorporates SPPFF (Spatial Pyramid Pooling Fusion Module) and C2PSA (Cross-Stage Partial Spatial Attention Module). C2PSA integrates a multi-head attention mechanism, enhancing multi-scale target detection in complex scenes. A Dynamic Label Assignment strategy adaptively adjusts the positive-to-negative sample ratio, improving recall for small target detection. The network architecture is illustrated in Figure 1.

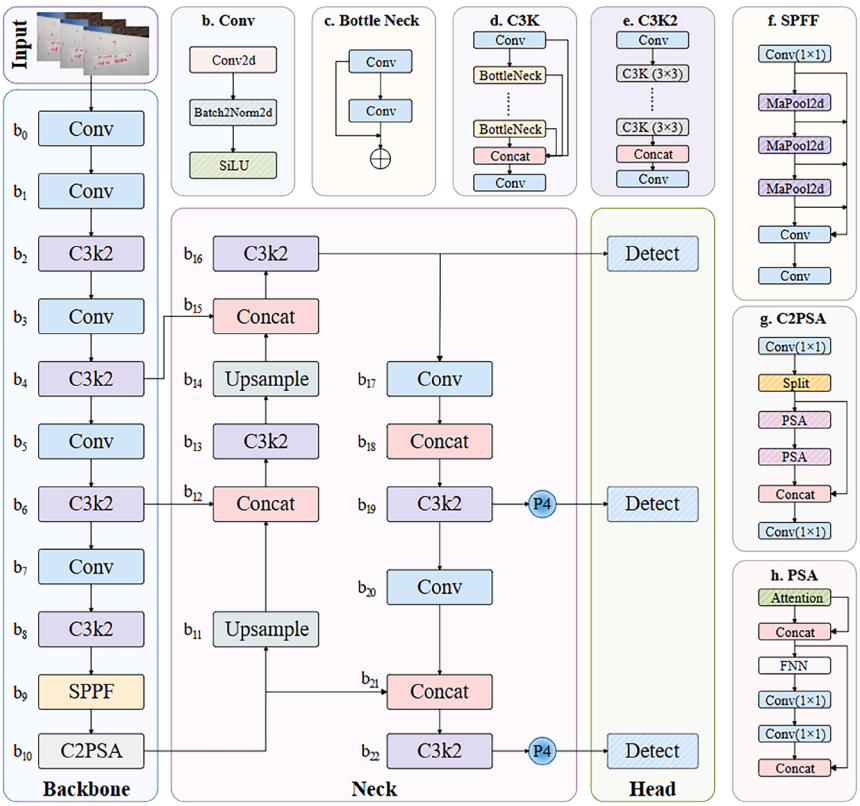

Fig. 1. YOLOv11 Network Architecture

2. ABS-YOLO. This paper presents ABS-YOLO, an enhanced model based on YOLOv11n baseline, to address object detection challenges in remote sensing imagery–particularly missed/false detections in complex backgrounds, dense scenes, and small targets. While maintaining a lightweight architecture, ABS-YOLO significantly improves detection accuracy and robustness through three key innovations:

AConv Module Integration. Deployed at the first backbone layer, this improved convolutional operation balances computational efficiency and feature representation via average pooling prior to convolution. By decomposing and enhancing traditional convolutions, it reduces computational complexity while improving key feature capture–critical for processing high-resolution remote sensing images and supporting small-target detection.

BiFPN Replacement. Substituting original Concat modules at layers 12 and 15, BiFPN employs bidirectional (top-down/bottom-up) pathways to fuse multi-scale features. Learnable weights dynamically prioritize feature importance, optimizing fusion capability.

Swin Transformer Integration. Added as the 23rd layer, this architecture leverages windowed self-attention for hierarchical feature extraction. It partitions images into non-overlapping windows for local self-attention with cross-window connections, enhancing multi-scale perception, especially for small targets, while controlling computational overhead. The improved model's network architecture is shown in Figure 2.

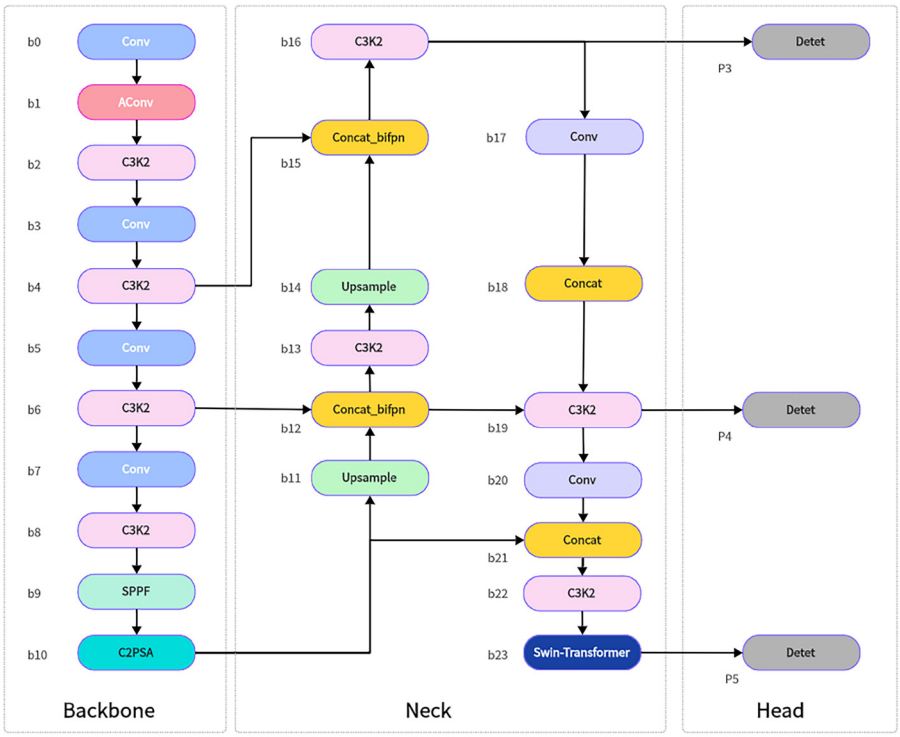

Fig. 2. ABS-YOLO Network Architecture

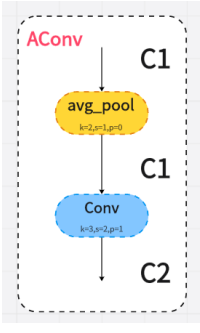

Fig. 3. AConv Structure

2.1. AConv. AConv module is a component of YOLO model. It mainly consists of a convolutional layer cv1, which has c1 input channels, c2 output channels, a kernel size of 3, a stride of 2, and padding of 1. This configuration enables the convolutional operation to downsample the input while extracting features. Prior to the convolutional operation, avg_pool2d function is applied to the input feature map for average pooling with a window size of 2 and a stride of 1. This operation reduces the size of the feature map while retaining important feature information. The specific structure of AConv is shown in Figure 3.

Compared to Conv module, AConv module adds an average pooling operation (*avg_pool*) on the basis of the basic convolutional layer. This design not only performs feature extraction but also downsamples the input feature map through the ave-

rage pooling operation, reducing the size of the feature map by half while retaining important feature information. The implementation of AConv module can be divided into two steps: average pooling and convolution. Combining the pooling and convolution operations, the overall expression of AConv module can be represented as:

$$y = CV_1\left(avg\_pool2d(x,2,1)\right),$$

where x is the input feature map; avg_pool2d $(x,2,1)$ indicates performing a 2 × 2 average pooling on $\boldsymbol{x}$ with a stride of 1; CV_1 is the convolutional layer with a kernel size of 3 × 3 a stride of 2, and padding of 1. The function of AConv module can be divided into two steps: performing 2 × 2 average pooling on the input feature map with a stride of 1; Conducting a 3 × 3 convolution operation on the pooled feature map with a stride of 2 and padding of 1. Through these two steps, the AConv module achieves downsampling and feature extraction of the feature map.

2.2. BiFPN. To address the detection needs of multi-scale targets in remote sensing images, such as small-sized vehicles and large-scale sports fields, ABS-YOLO model introduces BiFPN. BiFPN (Bidirectional Feature Pyramid Network) [18] is an efficient multi-scale feature fusion structure. Its core objective is to optimize the fusion of features across different levels through bidirectional cross-scale connections and dynamic weighting mechanisms, thereby enhancing detection accuracy especially for small targets and complex scenes in object detection tasks.

Conventional feature pyramid networks (e. g. FPN) enhance model performance through layer-wise consistency supervision, maintaining feature coherence across scales. However, FPN's [19] unidirectional top-down propagation (Figure 4, *a*) limits low-level feature utilization in final layers. PANet [20] introduces a bottom-up path (Figure 4, *b*) to augment low-level feature integration, while NAS-FPN [21] enhances FPN via additional layers (Figure 4, *c*), yet suffers from imbalanced feature contribution. BiFPN addresses these limitations through three key innovations (Figure 4, *d*), Node Pruning: removal of single-input nodes eliminates non-feature-fusion pathways; Skip Connections: direct links between same-level inputs/outputs enhance feature fusion without significant cost overhead; Layer Stacking: multiple bidirectional stages enable progressive feature refinement.

The core innovation of BiFPN lies in its weighted bi-directional feature fusion mechanism, which dynamically adjusts the contribution of different resolution features by introducing learnable weighting parameters. Its core formula contains the following two parts:

Fast Normalised Fusion (FNF). The formula is as follows:

$$O = \frac{\sum_i w_i \cdot l_i}{\varepsilon + \sum_i w_i},$$

where w_i is a learnable weight parameter, which ensures non-negativity through ReLU activation to avoid interference from negative weights; l_i is the input multi-scale features; ε is a minimal value to prevent division by zero. Through the learnable weight w_i, the network can adaptively distinguish the importance of features from different levels.

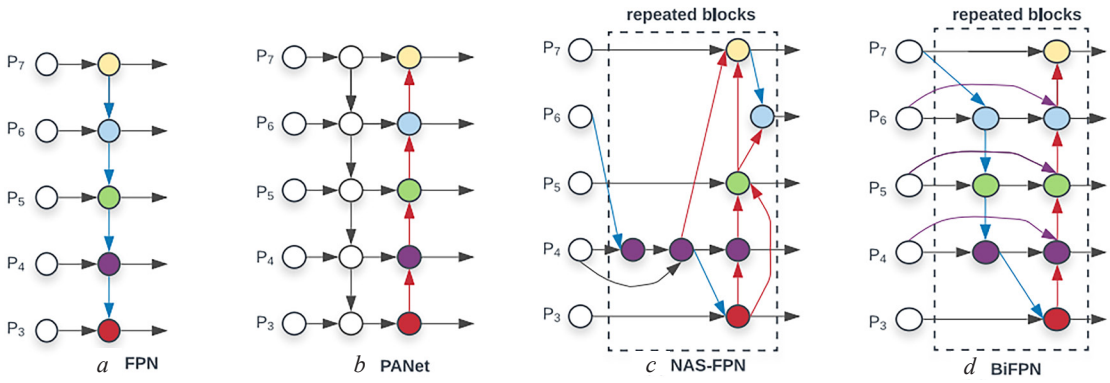

Fig. 4. Feature Network Design

Feature fusion through bidirectional paths. BiFPN achieves cross-scale feature interaction through top-down and bottom-up bidirectional paths. Take the output of a certain level l in the top-down path as an example.

The formula for the top-down path is:

$$P_l^{\text{td}} = \text{Conv}\left(\frac{w_1 \cdot P_l^{\text{in}} + w_2 \cdot \text{Resize}\left(P_{l+1}^{\text{in}}\right)}{w_1 + w_2 + \varepsilon}\right).$$

The formula for the bottom-up path is:

$$P_l^{\text{out}} = \text{Conv}\left(\frac{w_1' \cdot P_l^{\text{in}} + w_2' \cdot P_l^{\text{td}} + w_3' \cdot \text{Resize}\left(P_{l-1}^{\text{out}}\right)}{w_1' + w_1' + w_1' + \varepsilon}\right),$$

where P_l^{in} is the original input feature, and Resize is an upsampling or downsampling operation used for resolution alignment; w_1', w_2', w_3' are the weight parameters for different input paths. The top-down path conveys high-level semantic features to lower levels, providing semantic guidance for detail representation; the bottom-up path, in turn, strengthens the spatial localization capability of low-level features. The output of each node fuses the same-level input (P_l^{in}), the features passed from the upper layer $(P_{l+1}^{\text{td}}$ or $P_{l+1}^{\text{out}})$, and residual connection information, forming a closed-loop feature enhancement.

2.3. Swin Transformer. In remote sensing object detection tasks, high-resolution images in datasets contain a large number of small targets and complex background interferences. Traditional Convolutional Neural Networks (CNNs) are limited by their receptive fields and struggle to capture long-range contextual dependencies. Meanwhile, the global self-attention mechanism of the standard Vision Transformer (ViT) lacks hierarchical feature representation capabilities. To address these issues, this study integrates Swin Transformer to improve the attention mechanism of YOLOv11. Swin Transformer [22] is a visual model based on the Transformer architecture, proposed by researchers at Microsoft Research in 2021. Its full name is "Shifted Window Transformer". This model employs a novel window partitioning strategy to process images, enabling the Transformer to be more effectively applied to computer vision tasks. Swin Transformer adopts a hierarchical structure similar to CNNs, increasing the receptive field by reducing resolution layer by layer while reducing computational load. This makes it capable of effective handling high-resolution images. The core architecture of Swin Transformer is shown in the Figure 5.

The core advantages of Swin Transformer include. Hierarchical Feature Pyramid: by merging image patches (Patch Merging), it generates multi-scale feature maps (H/4 → H/32), which are well-suited for remote sensing scenes with significant variations in target sizes. Local Window Self-Attention: the input image is divided into non-overlapping windows (e. g., 7×7), and self-attention is computed within each window. This reduces the computational complexity to a linear level (O(n)), supporting high-resolution inputs. Window Shift Mechanism: in adjacent Transformer layers, the window positions are shifted to enable cross-window information interaction, enhancing the localization capability for small targets.

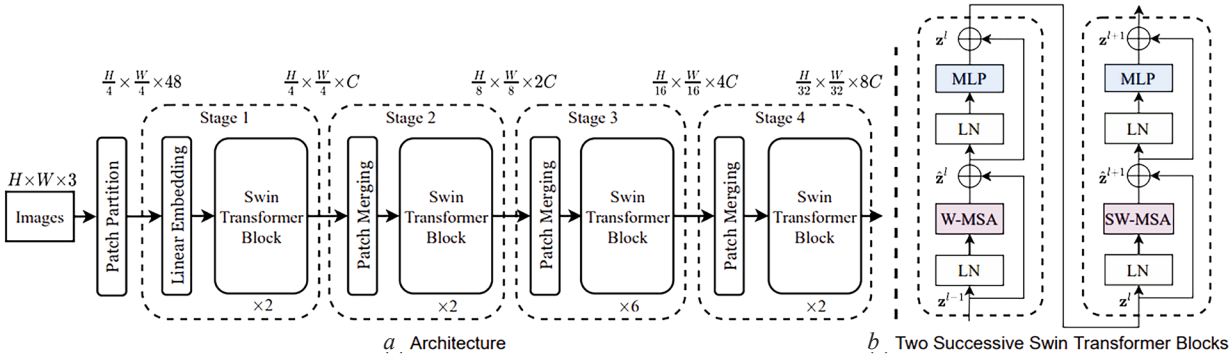

Fig. 5. Core Architecture of Swin Transformer

The window self-attention computation in Swin Transformer is as follows:

$$\text{Attention}\left(Q,K,V\right)=\text{SoftMax}\left(QK^{T}/\sqrt{d}+B\right)V,$$

where Q, K, and V are the query, key, and value matrices, respectively; d is the dimension; B is the relative position bias encoding, which is used to model the spatial relationships between pixels within a window.

The hierarchical window self-attention mechanism of Swin Transformer reduces the computational complexity of high-resolution images by computing within local windows (e.g., 7 ×7), while the window shifting strategy enables cross-region context modeling. This enhances the feature capture capability for small and densely packed targets. Its multi-scale pyramid structure (feature maps from H/4 to H/32), combined with BiFPN, efficiently fuses shallow details (such as ship edges) with deep semantic information (such as airport layouts). This addresses the issues of large target size variations and complex backgrounds in remote sensing scenes, ultimately reducing false detections while improving the recall rate and localization accuracy of small target detection.

3. Experimental Results and Analysis.

3.1. Experimental Environment and Datasets. All experiments in this study were completed in a unified hardware and software environment to ensure the reliability of the experimental results and the accuracy of the data. The specific environment configuration parameters of the experiment are shown in Table 1 below. The parameters not provided in this article use the official default parameters of YOLOv11n.

Table 1. **Experimental Parameter Settings**

Environment		Parameter	
Operating System	Windows 11 64-bit	Learning Rate	0.01
GPU	NVIDIA GeForce RTX 4060	Iterations	300
Memory	16G	Batchsize	16
Python	Python 3.9	Workers	0
Framework	PyTorch 2.4.0	Image Input Size	640×640
Environment	CUDA 12.41	Optimizer	Auto

The dataset used in the experiments is NWPU VHR-10 dataset [23, 24]. NWPU VHR-10 dataset is a challenging ten-category geospatial object detection dataset. It contains a total of 800 very high-resolution (VHR) optical remote sensing images, 715 color images of which were obtained from Google Earth with spatial resolutions ranging from 0.5 to 2 meters. Additionally, 85 sharpened color infrared images were acquired from the Vaihingen dataset with a spatial resolution of 0.08 meters. The dataset is divided into two groups: a) the positive image set, which contains at least one target in the image, consists of 650 images; b) the negative image set, which contains 150 images and does not include any targets. Consequently, the positive image set includes 757 airplanes, 302 ships, 655 storage tanks, 390 baseball diamonds, 524 tennis courts, 159 basketball courts, 163 ground tracks, 224 harbors, 124 bridges and 477 vehicles, all manually annotated with bounding boxes and used for ground truth instances.

3.2. Cross-Model Comparison Experiments. To validate the performance advantages of the proposed ABS-YOLO model in the remote sensing image object detection, we compared it with mainstream lightweight versions of YOLO series, including YOLOv5n [25], YOLOv6n [26], YOLOv8n [27], YOLOv10n [28], and YOLOv11n. The experimental results are shown in Table 2.

Through the experiments, we found that mAP50-95 of ABS-YOLO model is 0.551, which is a 2.6 % improvement over the baseline model's 0.537 and significantly better than YOLOv5n by + 3.8 % and YOLOv8n by 1.1 %. Although slightly lower than YOLOv6n's 0.557, the parameter count is reduced by 29.8 %, indicating that ABS-YOLO model has a greater advantage in parameter efficiency. mAP50 of the improved model reaches 0.901, surpassing all comparison models, which validates its precision in target localization and demonstrates stronger bounding box regression capabilities for small targets such

as airplanes and ships. The inference time of the improved model is 1.8 ms, which, although slightly slower than YOLOv8n (1.4 ms), still meets the real-time requirement (>30 FPS) and is suitable for offline high-precision detection scenarios. Figure 6 shows the detection effect of the algorithm before and after improvement on complex scene images in NWPU VHR-10 dataset. As it can be seen from the figure, the improved ABS-YOLO algorithm effectively reduces the problem of false detection of targets in complex scenes, and also verifies that ABS-YOLO has good accuracy.

Table 2. **Comparison of Different Models on NWPU VHR-10 Dataset**

Parameters / Models	YOLO11n	YOLOv10n	YOLOv8n	YOLOv6n	YOLOv5n	ABS-YOLO
P	0.911	0.883	0.9	0.932	0.889	0.932
R	0.812	0.772	0.826	0.804	0.825	0.84
mAP50	0.881	0.858	0.886	0.892	0.892	0.901
mAP50-95	0.537	0.522	0.545	0.557	0.531	0.551
Parameters / (million)	2.46	2.57	2.56	3.96	2.08	2.78
GFLOPs	6.3	8.2	6.8	11.5	5.8	16.6
Preprocess / (ms)	0.1	0.1	0.1	0.1	0.1	0.3
Inference / (ms)	1.5	1.5	1.4	1.7	1.6	1.8
Postprocess / (ms)	0.6	0.1	0.5	0.6	0.6	0.6

a *b* *c*

Fig. 6. Comparison diagram of NWPU-VHR-10 remote sensing dataset: *a* is the original image; *b* is YOLOv11n; *c* is ABS-YOLO

3.3. Comparison of Different Attention Mechanisms and Analysis of Results. To demonstrate the superiority of the proposed algorithm over some currently popular modules, a set of comparative experiments was conducted under the same experimental environment and parameters. The experimental results are shown in Table 3.

Table 3. **Comparison of Different Attention Mechanisms on NWPU VHR-10 Dataset**

Parameters / Models	YOLO11nbaseline	SimAM	SE	CAFMAttention	EMAAttention	DASI	ABS-YOLO
P	0.911	0.872	**0.912**	0.885	0.89	0.899	**0.932**
R	0.812	**0.856**	0.803	**0.823**	**0.844**	**0.841**	**0.84**
mAP50	0.881	**0.897**	**0.896**	**0.887**	**0.902**	**0.894**	**0.901**
mAP50-95	0.537	**0.546**	**0.545**	**0.543**	**0.549**	0.544	0.551
Parameters / (million)	2.46	2.46	2.46	2.8	2.47	3	2.78
GFLOPs	6.3	6.3	6.3	6.6	6.3	6.5	16.6
Preprocess / (ms)	0.1	0.1	0.1	0.1	0.1	0.1	0.3
Inference / (ms)	1.5	1.7	1.7	1.6	1.5	1.6	1.8
Postprocess / (ms)	0.6	0.5	0.5	0.6	0.6	0.7	0.6

EMA Attention [29] is an efficient multi-scale attention mechanism that balances accuracy and efficiency through local feature focusing and lightweight design, making it suitable for real-time edge detection. In this experiment, although the parameter count only increased by 0.04 % (258.5 thousand), and the inference speed remained consistent with the baseline at 1.5 ms, mAP50-95 improved by 2.2 %, especially with a significant increase in the recall rate for small targets to 0.844. SimAM [30] is a parameter-free attention mechanism that enhances feature representation by normalizing the feature map, without adding any parameters. In this experiment, SimAM maintained the baseline parameter count, achieved the highest recall rate of 0.856, and improved both mAP50 and mAP50-95. However, P value decreased slightly. CAFM Attention [31] is a cross-scale fusion attention mechanism that optimizes multi-scale features through multi-branch interaction. In this experiment, mAP50 and mAP50-95 improved by 1.6 and 1.5 %, respectively, but the recall rate was relatively low, and P value decreased. DASI [32] is a dynamic adaptive spatial attention mechanism that adjusts weights based on the input content. In this experiment, DASI achieved a recall rate improvement to 0.841.SE [33] is a channel attention module that enhances model performance by strengthening the overall channel features. In this experiment, mAP50 and mAP50-95 were increased by 1.7 and 1.5 % respectively. In summary, the above different attention mechanisms did not surpass ABS-YOLO in the results of mAP50 and mAP50-95.

3.4. Ablation Study Analysis. To verify the effectiveness of the proposed modules (AConv, BiFPN, Swin Transformer) for remote sensing image object detection tasks, we incrementally introduced different modules on YOLOv11n baseline model and designed the following ablation study variants:YOLOv11n – the baseline model without any improvement modules; YOLOv11n + AConv – introducing the adaptive convolution AConv in the backbone; YOLOv11n + BiFPN – replacing the original connections in the Neck part with BiFPN; YOLOv11n + Swin Transformer – introducing the Swin Transformer dynamic adaptive fusion module in the Neck part; ABS-YOLO – jointly using the AConv, BiFPN and Swin Transformer modules. The experimental results are shown in Table 4.

Table 4. **Ablation Study Results on NWPU VHR-10 Dataset**

Parameters / Models	YOLO11n	ACONV	BiFPN	SwinTransformer	ABS-YOLO
P	0.911	0.913	0.924	0.921	**0.932**
R	0.812	0.823	0.834	0.827	**0.84**
mAP50	0.881	0.893	**0.905**	0.893	**0.901**
mAP50-95	0.537	0.535	0.547	0.542	**0.551**
Parameters / (million)	2.46	2.46	2.46	2.78	2.78
GFLOPs	6.3	6.3	6.3	16.6	16.6
Preprocess / (ms)	0.1	0.1	0.1	0.1	0.3
Inference / (ms)	1.5	1.8	1.6	1.6	1.8
Postprocess / (ms)	0.6	0.6	**0.5**	0.6	0.6

In this experiment, AConv module slightly decreased mAP50-95 but increased the recall rate to 0.823 (+1.1 %). This indicates that AConv enhances local feature extraction by dynamically adjusting the convolutional kernel, but when used alone, it may introduce redundant computations, leading to a slight drop in accuracy. BiFPN significantly improved mAP50-95 to 0.547 (+1.0 %) and increased the recall rate to 0.834 (+2.2 %), demonstrating that the bidirectional weighted fusion mechanism effectively integrates multi-scale features, especially enhancing the detection capability for small targets. Swin Transformer achieved mAP50-95 of 0.542 (+0.5 %), but the parameter count surged to 2.91 million (+12.8 %) and the computational load increased to 16.6 GFLOPs (+163 %). This shows that the global attention modeling capability comes at a high computational cost, and using it alone is not cost-effective. The combined model ABS-YOLO reached mAP50 of 0.901, a 2.3 % improvement over the baseline, and mAP50-95 of 0.551, a 2.6 % improvement over the baseline, with a recall rate of 0.840, a 3.5 % increase over the baseline. However, due to the integration of Swin Transformer module, GFLOPs increased to 16.6. The model maximizes accuracy through the synergy of the modules. The experiments demonstrate the necessity of the synergy among the three modules; the effect of a single module is limited, but their combination significantly optimizes accuracy (ΔmAP = +2.6 %). In conclusion, the proposed model is better suited for object detection tasks in remote sensing data.

Conclusion. Aiming at the issue of limited detection accuracy in remote sensing images due to large target scale variations and numerous background interferences, this paper proposes a remote sensing image object detection model based on improved YOLOv11 – ABS-YOLO. By integrating adaptive convolution (AConv) into YOLOv11 model to replace standard convolutions in the shallow network, the initial feature extraction capability is enhanced. The bidirectional cross-scale connections (top-down + bottom-up) and weighted feature fusion mechanism optimize the transmission of multi-scale features, reducing the information loss associated with traditional FPN. The addition of Swin Transformer module enables hierarchical feature extraction to capture global context information, effective modeling large-scale spatial dependencies and enhancing local receptive fields, thereby significantly improving detection performance in complex remote sensing scenes. Experiments show that this method increases mAP50-95 to 0.551 on NWPU VHR-10 dataset, an absolute improvement of 2.6 % over the baseline model, with precision (0.932) and recall (0.84) both reaching optimal levels, thus verifying its effectiveness in addressing the core challenges of remote sensing detection. Although the single-image inference time of ABS-YOLO model is 1.8 ms, meeting real-time requirements and demonstrating the feasibility of prioritizing accuracy in remote sensing detection tasks, there is still room for optimization in terms of GFLOPs (16.6) and parameter scale (2.91 million). This study indicates that the synergistic design of local feature optimization (AConv), multi-scale fusion (BiFPN), and global modeling (Swin Transformer) can effectively break through the performance bottleneck of traditional YOLO series models in remote sensing image detection, providing a high-precision and high-efficiency solution for high-resolution remote sensing target detection. Future work will investigate the engineering feasibility of dynamic computational pruning and multi-modal fusion, as well as explore the collaborative optimization path of NAS and edge deployment, to advance remote sensing detection technology towards real-time and intelligent directions.

References

1. Velastegui-Montoya A., Montalván-Burbano N., Carrión-Mero P., Rivera-Torres H., Sadeck L., Adami M. Google Earth Engine: A Global Analysis and Future Trends. *Remote Sensing*, 2023, vol. 15, no. 14, art. ID 3675. https://doi.org/10.3390/rs15143675

2. Liu H., Gong P., Wang J., Wang X., Ning G., Xu B. Production of global daily seamless data cubes and quantification of global land cover change from 1985 to 2020 – iMap World 1.0. *Remote Sensing of Environment*, 2021, vol. 258, art. ID 112364. https://doi.org/10.1016/j.rse.2021.112364

3. Zhang L. P., Shen H. F. Progress and future of remote sensing data fusion. *Journal of Remote Sensing*, 2016, vol. 20, no. 5, pp. 1050–1061. https://doi.org/10.11834/jrs.20166243

4. Dalal N., Triggs B. Histograms of oriented gradients for human detection. *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2005, vol. 1, pp. 886–893. https://doi.org/10.1109/CVPR.2005.177

5. Lowe D. G. Object recognition from local scale-invariant features. *Proceedings of the Seventh IEEE International Conference on Computer Vision*, 1999, vol. 2, pp. 1150–1157. https://doi.org/10.1109/ICCV.1999.790410

6. Maturana D., Mery D., Soto Á. Face recognition with local binary patterns, spatial pyramid histograms and naive Bayes nearest neighbor classification. *Proceedings of the 2009 International Conference of the Chilean Computer Science Society*, 2009, pp. 125–132. https://doi.org/10.1109/SCCC.2009.21

7. Girshick R., Donahue J., Darrell T., Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. *Arxiv* [Preprint], 2014. Available at: https://arxiv.org/abs/1311.2524. https://doi.org/10.48550/arXiv.1311.2524

8. Girshick R. Fast R-CNN. *Arxiv* [Preprint], 2015. Available at: https://arxiv.org/abs/1504.08083. https://doi.org/10.48550/arXiv.1504.08083

9. Ren S., He K., Girshick R., Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. *Arxiv* [Preprint], 2016. Available at: https://arxiv.org/abs/1506.01497. https://doi.org/10.48550/arXiv.1506.01497

10. Redmon J., Divvala S., Girshick R., Farhadi A. You only look once: Unified, real-time object detection. *Arxiv* [Preprint], 2016. Available at: https://arxiv.org/abs/1506.02640. https://doi.org/10.48550/arXiv.1506.02640

11. Liu W., Anguelov D., Erhan D., Szegedy C., Reed S., Fu C.-Y., Berg A. C. SSD: Single Shot MultiBox Detector. *Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science, vol 9905*. Springer, 2016, pp. 21–37. https://doi.org/10.1007/978-3-319-46448-0_2

12. Zhang Y., Ye M., Zhu G., Liu Y., Guo P., Yan J. FFCA-YOLO for small object detection in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 2024, vol. 62, pp. 1–15. https://doi.org/10.1109/TGRS.2024.3363057

13. Yi H., Liu B., Zhao B., Liu E. Small object detection algorithm based on improved YOLOv8 for remote sensing. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2024, vol. 17, pp. 1734–1747. https://doi.org/10.1109/JSTARS.2023.3339235

14. Wu T., Dong Y. YOLO-SE: Improved YOLOv8 for Remote Sensing Object Detection and Recognition. *Applied Sciences*, 2023, vol. 13, no. 24, art. ID 12977. https://doi.org/10.3390/app132412977

15. Wang X., Gao H., Jia Z., Li Z. BL-YOLOv8: An Improved Road Defect Detection Model Based on YOLOv8. *Sensors*, 2023, vol. 23, no. 20, art. ID 8361. https://doi.org/10.3390/s23208361

16. Soudeep S., Jahin M. A., Mridha M. F. Interpretable dynamic graph neural networks for small occluded object detection and tracking. *Arxiv* [Preprint], 2025. Available at: https://arxiv.org/abs/2411.17251. https://doi.org/10.48550/arXiv.2411.17251

17. Khanam R., Hussain M. YOLOv11: An overview of the key architectural enhancements. *Arxiv* [Preprint], 2024. Available at: https://arxiv.org/abs/2410.17725. https://doi.org/10.48550/arXiv.2410.17725

18. Tan M., Pang R., Le Q. V. EfficientDet: Scalable and efficient object detection. *Arxiv* [Preprint], 2020. Available at: https://arxiv.org/abs/1911.09070. https://doi.org/10.48550/arXiv.1911.09070

19. Lin T.-Y., Dollár P., Girshick R., He K., Hariharan B., Belongie S. Feature pyramid networks for object detection. *Arxiv* [Preprint], 2017. Available at: https://arxiv.org/abs/1612.03144. https://doi.org/10.48550/arXiv.1612.03144

20. Liu S., Qi L., Qin H., Shi J., Jia J. Path aggregation network for instance segmentation. *Arxiv* [Preprint], 2018. Available at: https://arxiv.org/abs/1803.01534. https://doi.org/10.48550/arXiv.1803.01534

21. Ghiasi G., Lin T.-Y., Pang R., Le Q. V. NAS-FPN: Learning scalable feature pyramid architecture for object detection. *Arxiv* [Preprint], 2019. Available at: https://arxiv.org/abs/1904.07392. https://doi.org/10.48550/arXiv.1904.07392

22. Liu Z., Lin Y., Cao Y., Hu H., Wei Y., Zhang Z., Lin S., Guo B. Swin Transformer: Hierarchical vision transformer using shifted windows. *Arxiv* [Preprint], 2021. Available at: https://arxiv.org/abs/2103.14030. https://doi.org/10.48550/arXiv.2103.14030

23. Su H., Wei S., Yan M., Wang C., Shi J., Zhang X. Object detection and instance segmentation in remote sensing imagery based on precise Mask R-CNN. *IGARSS 2019 – 2019 IEEE International Geoscience and Remote Sensing Symposium*, 2019, pp. 1454–1457. https://doi.org/10.1109/IGARSS.2019.8898573

24. Su H., Wei S., Liu S., Liang J., Wang C., Shi J., Zhang X. HQ-ISNet: High-Quality Instance Segmentation for Remote Sensing Imagery. *Remote Sensing*, 2020, vol. 12, no. 6, pp. 989. https://doi.org/10.3390/rs12060989

25. Khanam R., Hussain M. What is YOLOv5: A deep look into the internal features of the popular object detector. *Arxiv* [Preprint], 2024. Available at: https://arxiv.org/abs/2407.20892. https://doi.org/10.48550/arXiv.2407.20892

26. Li C., Li L., Jiang H., Weng K., Geng Y., Li L., Ke Z. [et al.]. YOLOv6: A single-stage object detection framework for industrial applications. *Arxiv* [Preprint], 2022. Available at: https://arxiv.org/abs/2209.02976. https://doi.org/10.48550/arXiv.2209.02976

27. Jocher G., Qiu J., Chaurasia A. Ultralytics YOLO (Version 8.0.0) [Computer software]. 2023. Available at: https://github.com/ultralytics/ultralytics

28. Wang A., Chen H., Liu L., Chen K., Lin Z., Han J., Ding G. YOLOv10: Real-time end-to-end object detection. *Arxiv* [Preprint], 2024. Available at: https://arxiv.org/abs/2405.14458. https://doi.org/10.48550/arXiv.2405.14458

29. Ouyang D., He S., Zhang G., Luo M., Guo H., Zhan J., Huang Z. Efficient multi-scale attention module with cross-spatial learning. *ICASSP 2023 – 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2023, pp. 1–5. https://doi.org/10.1109/ICASSP49357.2023.10096516

30. Yang L., Zhang R.-Y., Li L., Xie X. SimAM: A simple, parameter-free attention module for convolutional neural networks. *Proceedings of the 38th International Conference on Machine Learning*, *PMLR*, 2021, vol. 139, pp. 11863–11874. Available at: https://proceedings.mlr.press/v139/yang21o.html

31. Chen Z., Lu S. CAF-YOLO: A Robust framework for multi-scale lesion detection in biomedical imagery. *Arxiv* [Preprint], 2024. Available at: https://arxiv.org/abs/2408.01897. https://doi.org/10.48550/arXiv.2408.01897

32. Xu S., Zheng S., Xu W., Xu R., Wang C., Zhang J., Teng X., Li A., Guo L. HCF-Net: Hierarchical context fusion network for infrared small object detection. *Arxiv* [Preprint], 2024. Available at: https://arxiv.org/abs/2403.10778. https://doi.org/10.48550/arXiv.2403.10778

33. He K., Gkioxari G., Dollár P., Girshick R. Mask R-CNN. *Arxiv* [Preprint], 2018. Available at: https://arxiv.org/abs/1703.06870. https://doi.org/10.48550/arXiv.1703.06870

Information about the authors

Sergey V. Ablameyko – Academician of the National Academy of Sciences of Belarus, Dr. Sc. (Engineering), Professor, United Institute of Informatics Problems of the National Academy of Sciences of Belarus (6, Surganov Str., 220012, Minsk, Republic of Belarus); Belarusian State University (4, Nezavisimosti Ave., 220030, Minsk, Republic of Belarus). E-mail: ablameyko@yandex.by. https://orcid.org/0000-0001-9404-1206

Wu Xianyi – Postgraduate Student, Belarusian State University (4, Nezavisimosti Ave., 220030, Minsk, Republic of Belarus). E-mail: tigerv5872@gmail.com. https://orcid.org/0009-0003-6976-5386

Информация об авторах

Абламейко Сергей Владимирович – академик Национальной академии наук Беларуси, доктор технических наук, профессор, Объединенный институт проблем информатики Национальной академии наук Беларуси (ул. Сурганова, 6, 220012, Минск, Республика Беларусь); Белорусский государственный университет (пр. Независимости, 4, 220030, Минск, Республика Беларусь). E-mail: ablameyko@yandex.by. https://orcid.org/0000-0001-9404-1206

Ву Сяньи – аспирант, Белорусский государственный университет (пр. Независимости, 4, 220030, Минск, Республика Беларусь). E-mail: tigerv5872@gmail.com. https://orcid.org/0009-0003-6976-5386