UDC 004.93
DOI: 10.21122/2309-4923-2025-3-4-10

XIANGYI WU[1], SERGEY V. ABLAMEYKO[1,2]

REMOTE SENSING IMAGE TARGET DETECTION MODEL INTEGRATING DYNAMIC RECEPTIVE FIELD AND SNAKE CONVOLUTION

[1] Belarusian State University, Minsk, 220030, Republic of Belarus
[2] United Institute of Informatics Problems, National Academy of Sciences of Belarus, Minsk, Republic of Belarus

Abstract. *To address the challenges of a high missed detection rate for small targets and strong interference from complex backgrounds in remote sensing image target detection, the paper proposes an improved YOLOv11n based method. We introduce an enhanced YOLOv11n model incorporating a dynamic receptive field module (RFAConv) and a snake deformation modeling module (DySnakeConv). This approach strengthens shallow feature extraction capabilities and refines adaptive fitting of target boundaries, thereby improving detection accuracy. Experimental results demonstrate that on the RSOD dataset, the improved model achieves mean average precision (mAP) scores of 96.9 % at IoU = 0.50 (mAP50) and 65.5 % over IoU thresholds from 0.50 to 0.95 (mAP50-95). These results surpass those of YOLOv8n, YOLOv10n, and other comparative models in key metrics such as precision and recall. Importantly, the model maintains comparable performance on the NWPU VHR-10 dataset. The proposed model presents an efficient solution for detecting small and geometrically sensitive targets in high-resolution remote sensing images.*

Keywords: *Remote Sensing Image, Object Detection, YOLOv11, RFAConv, DySnakeConv*

1. Introduction

The core task of remote sensing image target detection lies in precisely locating and identifying specific targets within the image. This technology holds extensive applications across national defense, military, and national economic fields [1]. In remote sensing image detection tasks, the significant variations in target size and the high complexity of background conditions greatly increase the difficulty of detection and recognition. Contrasted with traditional natural images, remote sensing images feature more intricate backgrounds. The target information tends to be fragmented and densely distributed, causing the feature map to be filled with a vast amount of interference information, which in turn further intensifies the challenges of detection.

With the development of deep learning, deep learning-based remote sensing image target detection methods have achieved remarkable progress [2]. These methods effectively address complex scenarios in remote sensing imagery, improving both detection accuracy and real-time capability. Currently, deep learning-based target detection algorithms are primarily categorized into two distinct types: single-stage detection frameworks and two-stage detection frameworks. The two-stage detection paradigm is typified by the R-CNN [3] series, which encompasses enhanced variants such as Fast R-CNN and Faster R-CNN. In contrast, the single-stage detection paradigm is represented by mainstream algorithms including SSD [4], RetinaNet [5], and YOLO [6].

Although deep learning-based remote sensing target detection algorithms have achieved substantial advancements, they persistently confront challenges such as inadequate detection accuracy and low computational efficiency. To address these challenges, researchers have developed diverse optimization strategies. Betti et al. [7] introduced a YOLO variant optimized for small target detection, employing a compact feature extractor and skip connections to reuse features and integrate information, thereby enhancing detection performance. Wu et al. [8] enhanced the YOLO network by integrating an attention mechanism, augmenting feature fusion, and incorporating a dedicated small target detection layer, which elevated remote sensing image detection accuracy. Han et al. [9] employed dilated convolutions to capture multi-scale information, utilizing varying dilation rates to expand the receptive field. Nie et al. [10] proposed the SSFF module and utilized HPANet to replace the Path Aggregation Network, resulting in enhanced accuracy and reduced network parameters.

In response to the characteristics of remote sensing images, including complex backgrounds, diverse target shapes, and varied scales, this paper proposes an enhanced algorithm termed YOLO-RDC (YOLOv11+RepConv+DySnakeConv), which is based on YOLOv11, designed to optimize the model structure and feature fusion strategy, thereby mitigating issues such as small target missed detection and geometric target deformation in remote sensing target detection tasks.

2. The model structure

Figure 1 illustrates the architecture of the proposed YOLO-RDC network model. First, the Receptive Field Attention Conv (RFAConv) module is integrated

into the shallow backbone layer (P3/8) to enhance small-target features, utilizing multi-scale convolutions and attention weighting to amplify texture details on 80×80 high-resolution feature maps, thereby suppressing complex background interference. Subsequently, the C2f_DySnakeConv module is applied to the deep head layer (P5/32) for optimizing geometrically sensitive target boundaries, where deformable convolution kernels adaptively fit contours along target centerlines, addressing deformation challenges in irregular structures (e.g., runways, bridges, roads).

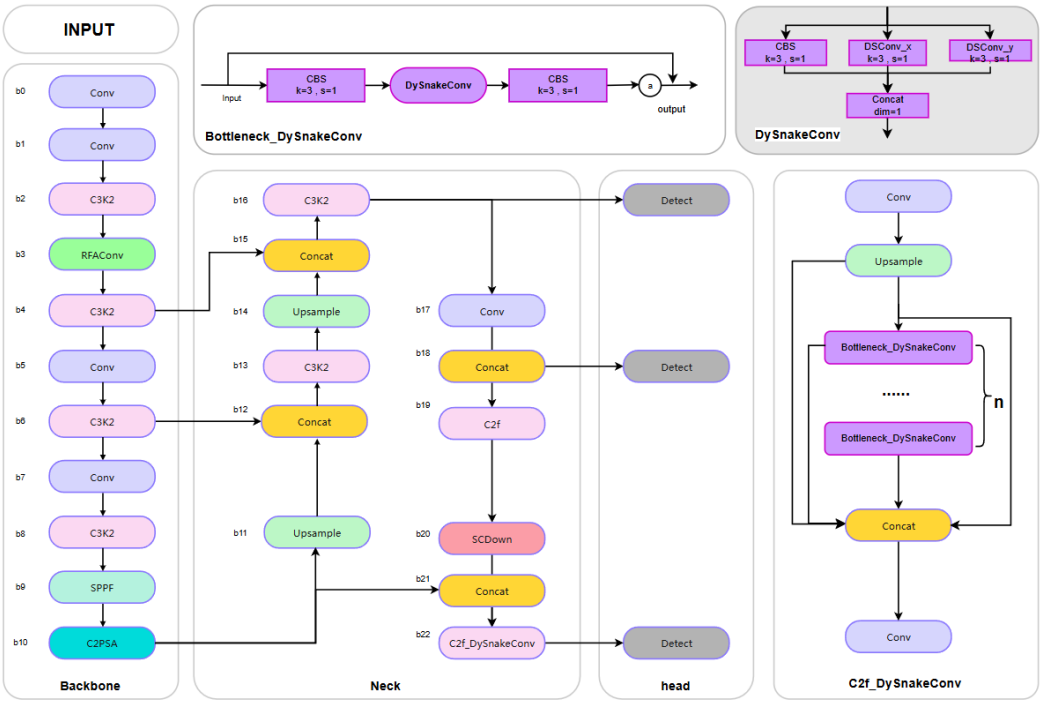


Figure 1. Improved model network structure diagram

Shallow features exhibit higher resolution yet weaker semantics, enabling detailed extraction; whereas deep features possess stronger semantics but lower resolution, facilitating structural modeling. Both feature hierarchies transmit geometric information via the Feature Pyramid Network (FPN)'s top-down pathway to preserve high-level feature details.

2.1 RFAConv

RFAConv (Receptive-Field Attention Convolution) integrates spatial attention mechanisms with convolutional operations to enhance convolutional neural networks (CNNs) [11]. By concentrating on spatial features within the receptive field, RFAConv enables CNNs to interpret local regions more effectively, thereby improving feature extraction accuracy. Standard spatial attention alone may be insufficient to capture critical information in large convolution kernels. RFAConv addresses this limitation by generating optimized attention weights, allowing large kernels to process image information more efficiently. The detailed architecture of RFAConv is illustrated in Figure 2.

RFAConv employs grouped convolution to extract spatial features of the receptive field and generate independent sliding windows, then utilizes global average pooling and a 1×1 convolutional layer to compute attention weights, and finally multiplies the features by these weights to produce the output. In summary, the RFA operation can be formulated as:

$$F = \mathrm{Softmax}\left(g^{1\times1}(\mathrm{AvgPool}(X))\right) \times \mathrm{ReLU}\left(\mathrm{Norm}\left(g^{k\times k}(X)\right)\right) = A_{rf} \times F_{rf} \quad (1)$$

AvgPool aggregates global information within each receptive field, while **1×1** group convolution processes cross-channel interactions. The ***Softmax*** operator highlights discriminative features across the receptive field. $\mathbf{g}^{i\times i}$ is group convolution with kernel size $i \times i$; k is convolution kernel size; ***Norm*** is normalization layer; X is input feature map; $\mathbf{A}_{rf}$ denotes the attention map and $\mathbf{F}_{rf}$ the spatial feature of the receptive field.

2.2 DySnakeConv

DySnakeConv (Dynamic Snake Convolution) is a dynamic convolution method designed to enhance model expressiveness and efficiency through adaptive kernel adjustment. This enables convolutional neural networks to accurately perceive and focus on slender, curved local structures [12]. These structures occupy a small pixel proportion in images and exhibit variable

shapes. Standard or traditional deformable convolutions struggle to capture these features stably and accurately, often resulting in segmentation breaks. Consequently, the core concept of DySnakeConv involves incorporating a deformable mechanism, allowing convolution operations to automatically adapt receptive field size and shape based on input image regions. The detailed architecture of DySnakeConv is illustrated in Figure 3.

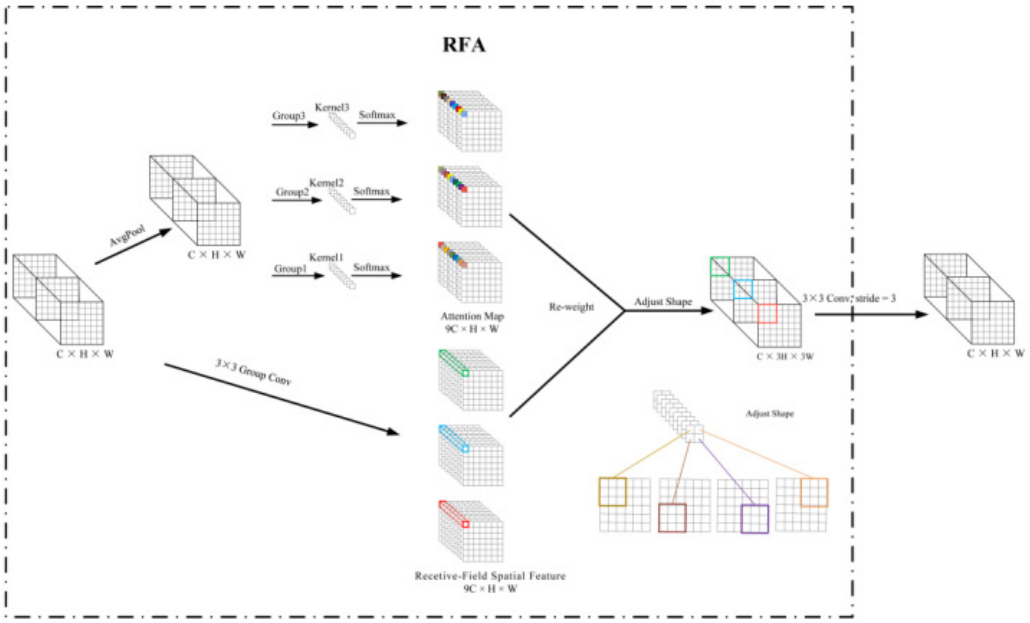


Figure 2. The detailed structure of RFAConv

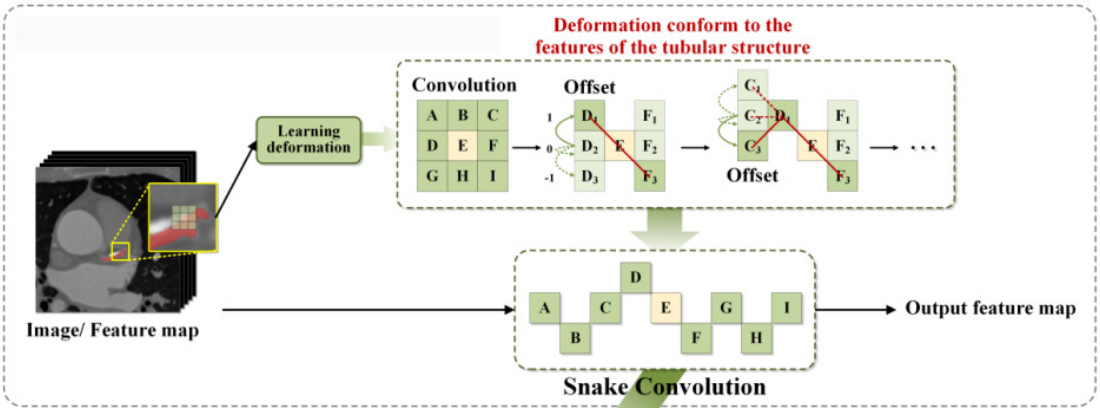


Figure 3. DySnakeConv structure diagram

In essence, DySnakeConv enables the convolution kernel's sampling points to initiate from the center and progressively shift along a primary direction. During each step, the network acquires minute vertical offsets and progressively accumulates them. The sampling points of the convolution kernel are connected into a continuous curved path, and then feature extraction is performed.Set the center coordinate of the convolution to $K_i = (x_i, y_i)$, Δ is the deformation offset, and c represents the horizontal distance from the center grid. The calculation of DySnakeConv can be expressed as:

$$\mathbf{K}_{i\pm c} = \left\{ \left(\mathbf{x}_{i+c}, \mathbf{y}_{i+c}\right) = \left(\mathbf{x}_i + \mathbf{c}, \mathbf{y}_i + \sum_i^{i+c} \text{"}\mathbf{y}\right) \right\},$$
$$\mathbf{K}_{j\pm c} = \left\{ \left(\mathbf{x}_{j+c}, \mathbf{y}_{j+c}\right) = \left(\mathbf{x}_j + \sum_j^{j+c} \text{"}\mathbf{y}, \mathbf{y}_j + \mathbf{c}\right) \right\}. \quad (2)$$
$$\left(\mathbf{x}_{j-c}, \mathbf{y}_{j-c}\right) = \left(\mathbf{x}_j - \sum_{j-c}^{j} \text{"}\mathbf{y}, \mathbf{y}_j - \mathbf{c}\right)$$

$K_{i\pm c}$ is the calculation in the **x**-axis direction, $K_{j\pm c}$ is the calculation in the **y**-axis direction, Σ is the accumulated offset, and since the offset Δ is usually a fractional value, bilinear interpolation needs to be implemented:

$$K = \sum_{K'} B(K', K) \cdot K', \quad (3)$$

where K represents the fractional coordinate position in formulas (2), K' enumerates all integer spatial positions, and **B** is the bilinear interpolation kernel. Due to the change in two dimensions (x-axis, y-axis), DySnakeConv covers a 9×9 range during the deformation process. DySnakeConv aims to better adapt to slender structures based on dynamic structures, so as to more effectively perceive key features.

3. Experiments and results analysis

3.1 Experimental environment and parameter settings

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 configuration table

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

3.2 Dataset and Evaluation Criteria

RSOD dataset [13] is a professional remote sensing image object detection dataset released by Wuhan University in 2015, including the following four types of targets: Airplane: 446 images in total, including 4993 aircraft. Playground: 189 images in total, including 191 playgrounds. Overpass: 176 images in total, including 180 overpasses. Oil drum: 165 images in total, including 1586 oil drums. NWPU VHR-10 dataset [14] is an open source dataset released by Northwestern Polytechnical University in 2014. It has 650 images containing targets and 150 background images, totaling 800 images.

3.3 Cross-model comparison experiments

In this experiment, multiple target detection algorithms are selected for performance comparison, including YOLOv5n, YOLOv6n, YOLOv8n and YOLOv10n. The comparison results are shown in Table 2:

Table 2. Comparison of different models on NWPU VHR-10 and RSOD datasets

NWPU VHR-10					RSOD			
model	P	R	mAP50	mAP50-95	P	R	mAP50	mAP50-95
YOLOv5n	88.9	82.5	89.2	53.1	94.8	89.4	94.3	63.2
YOLOv6n	93.2	80.4	89.2	55.7	93.1	89.9	94.4	64.3
YOLOv8n	90	82.6	88.6	54.5	91.6	89.1	94	64.9
YOLOv10n	88.3	77.2	85.8	52.2	90.5	88.4	94.3	64.7
YOLO11n	91.1	81.2	88.1	53.7	95.9	92.1	96.4	64.3
Ours	91.6	81.5	90	54.5	96	92.2	96.9	65.5

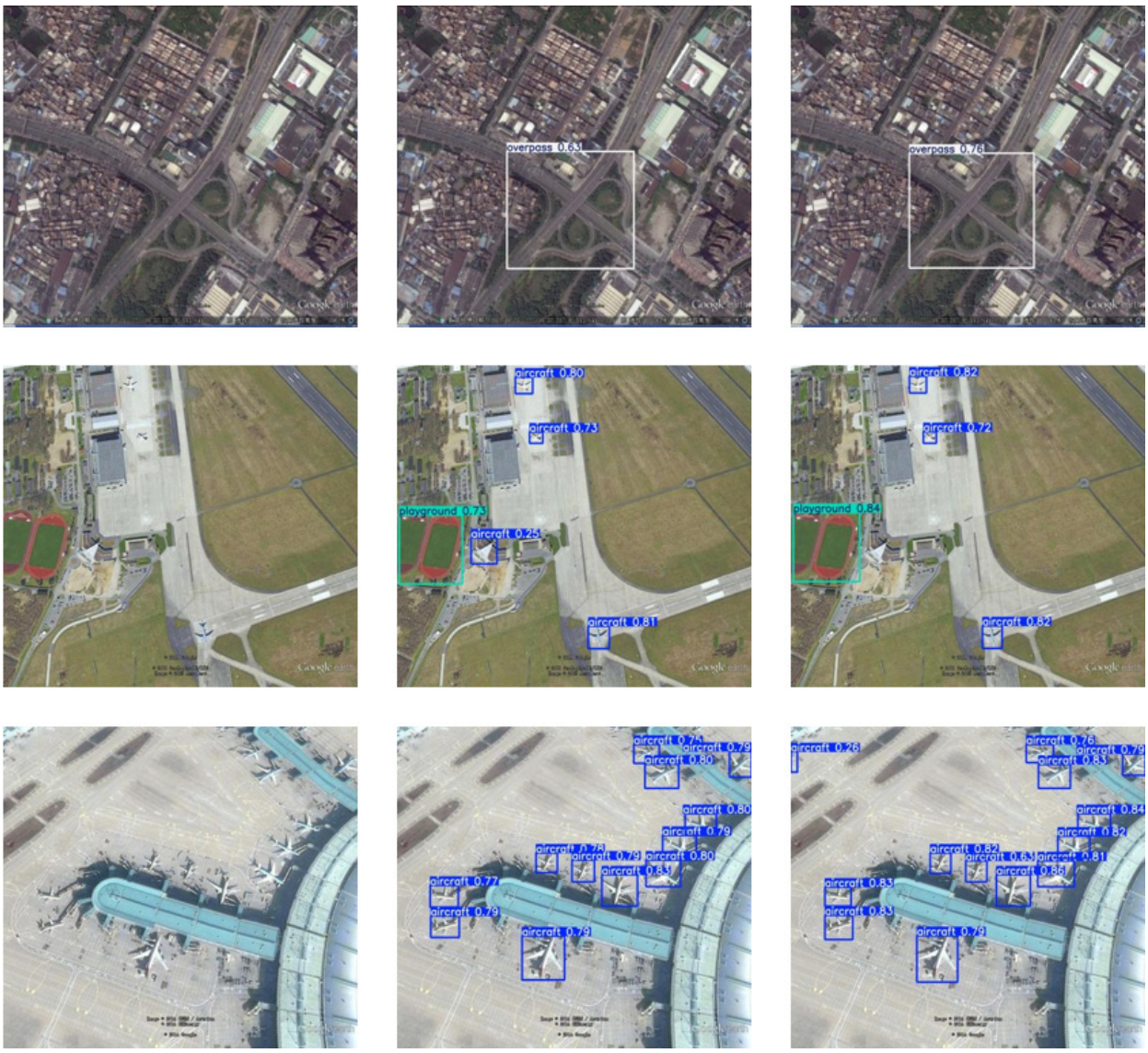
As shown in Table 2, the YOLO-RDC model performs well in the evaluation indicators on both datasets. Compared with YOLO11n, YOLO-RDC has an mAP50 that is 1.9 % higher and mAP50-95 that is 0.8 higher on the NWPU VHR-10 dataset, and an mAP50 that is 0.5 % higher and mAP50-95 that is 1.2 % higher on the RSOD dataset, demonstrating its recognition accuracy and reliability for multi-category targets in complex scenarios.

Table 3 shows the detection results of the model on all categories on the RSOD dataset.

Table 3. Comparison of experimental results for all categories (mAP50)

	YOLOv5n	YOLOv6n	YOLOv8n	YOLOv10n	YOLOv11n	Ours
all	94.3	94.4	94	94.3	96.4	96.9
aircraft	94.9	95.8	95.7	94.8	96.4	96.1
oiltank	96.4	96.5	96.7	96.2	95.7	97.5
overpass	86.2	85.7	84.2	86.5	93.9	94.6
playground	99.5	99.5	99.3	99.5	99.5	99.5

As shown in Table 3, this YOLO-RDC performs well in various target detection experiments, especially in oiltank and overpass. Figure 4 shows the detection effect of the algorithm before and after improvement on complex scene images in the ROSD dataset. As can be seen from the figure, the improved YOLO-RDC algorithm effectively improves the detection accuracy of curved targets, while also reducing the problem of false detection and missed detection of targets in complex scenes.



a. YOLO-RDC *b*. YOLO-RDC *c*. YOLO-RDC

Figure 4. Comparison of ROSD remote sensing datasets

In order to further verify the advantages of the algorithm in this paper, common attention modules such as EfficientNetv2, EMA Attention, RepViTblock, SE, Swin Transformer are introduced for comparative experiments. The experimental results are shown in Table 4.

Table 4. Experimental results comparing the different attention mechanisms

Model	P	R	mAP50	mAP50-95	parameters	GFLOPs
YOLOv11n	95.9	92.1	96.4	64.3	2.46M	6.3
EfficientNetv2	94.3	89.6	93.6	60.3	2M	3
EMA_attention	91.1	91.2	94.9	64.8	2.46M	6.3
RepViTblock	94.8	90.4	95.3	64.2	2.78M	6.3
SE	88	92.5	93.1	63.5	2.46M	6.3
SwinTransformer	93.2	91.9	94.5	63.8	2.78M	16.6
Ours	96	92.2	96.9	65.5	2.8M	6.6

The P, mAP50, and mAP50-95 of the proposed method are 96 %, 96.9 %, and 65.5 %, respectively, all reaching the highest values in the experiment. Although the model size and detection speed are not optimal, they are also within an acceptable range. This verifies the advantages of the proposed algorithm in remote sensing image detection.

4. Conclusion

In this paper, an improved model based on the YOLOv11n architecture is proposed to address the challenges of low accuracy in detecting small targets in remote sensing images and strong background interference. The local feature extraction capability is enhanced by introducing RFAConv, and the target boundary fitting is optimized by fusing the C2f_DySnakeConv structure, which significantly improves the detection robustness in complex scenes. Experimental results on the NWPU VHR-10 and RSOD datasets show that the model has achieved significant improvements in both accuracy and efficiency. The mAP50 of the model reaches 90.0 % and 96.9 % respectively, which is significantly improved compared with the baseline YOLOv11n, especially in difficult categories such as oil tanks and overpasses, which is improved by more than 10 %. In terms of efficiency, although the number of parameters has increased, it is acceptable compared with the increase in computational cost and the significant improvement in the average accuracy of the model. Future research directions will focus on heterogeneous modality fusion and edge deployment optimization to further improve the performance and applicability of the model.

REFERENCES

1. Li AB, Guo H, Qi C, et al. Dense object detection in remote sensing images under complex background. Computer Engineering and Applications. 2023;59(8):247–253.

2. Yan J, Hu X, Zhang K, Shi T, Zhu G, Zhang Y. Multi-level feature fusion based dim small ground target detection in remote sensing images. Chinese Journal of Scientific Instrument. 2022;4(3):221–229. https://library.imaging.org/jist/articles/67/1/010505

3. Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2017;39(6),1137–1149. **DOI:** 10.1109/TPAMI.2016.2577031

4. Jang Y, Gunes H, Patras I. Registration-free Face-SSD: Single shot analysis of smiles, facial attributes, and affect in the wild. Computer Vision and Image Understanding. 2019;182:17–29. **DOI:** 10.1016/j.cviu.2019.01.006

5. Lin T-Y, Goyal P, Girshick R, He K, Dollár P. Focal Loss for Dense Object Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2020;42(2):318–327. **DOI:** 10.1109/TPAMI.2018.2858826

6. Redmon J, Divvala S, Girshick R, Farhadi A. (2016). You Only Look Once: Unified, Real-Time Object Detection. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA; 2016. Pp. 779-788. **DOI:** 10.1109/CVPR.2016.91

7. Betti A, Tucci M. YOLO-S: A Lightweight and Accurate YOLO-like Network for Small Target Detection in Aerial Imagery. Sensors. 2023;23(4):1865. **DOI:** 10.3390/s23041865

8. Wu X, Ablameyko SV. Efficient detection of building in remote sensing images using an improved YOLOv10 network. Informatika [Informatics]. 2025;22(2):33–47. **DOI:** 10.37661/1816-0301-2025-22-2-33-47

9. Han W, Kuerban A, Yang Y, Huang Z, Liu B, Gao J. (2022). Multi-Vision Network for Accurate and Real-Time Small Object Detection in Optical Remote Sensing Images. IEEE Geoscience and Remote Sensing Letters. 2022;19:1–5. **DOI:** 10.1109/LGRS.2020.3044422

10. Nie H, Pang H, Ma M, Zheng R. A Lightweight Remote Sensing Small Target Image Detection Algorithm Based on Improved YOLOv8. Sensors (Basel). 2024;24(9):2952. **DOI:** 10.3390/s24092952

11. Wu Z, Geiger A., Rozner J, Kreiss E, Lu H, Icard T, Potts C, & Goodman ND. Training trajectories of language models across scales. 2023. arXiv preprint arXiv:2304.03198. https://arxiv.org/pdf/2304.03198

12. Mirchandani S, Xia F, Florence P, Ichter B, Driess D, Arenas MG, et al. Large Language Models as General Pattern Machines. Proceedings of the 7th Conference on Robot Learning (CoRL). Atlanta, USA; 2023. pp. 2498-2518. Available at: https://general-pattern-machines.github.io (accessed 02.07.2025). **DOI:** 10.48550/arXiv.2307.04721

13. Long Y, Gong Y, Xiao Z, Liu Q. Accurate Object Localization in Remote Sensing Images Based on Convolutional Neural Networks. IEEE Transactions on Geoscience and Remote Sensing. 2017;55(5):2486-2498. **DOI:** 10.1109/TGRS.2016.2645610

14. Cheng G, Han J, Zhou P, Guo L. Multi-class geospatial object detection and geographic image classification based on collection of part detectors. ISPRS Journal of Photogrammetry and Remote Sensing. 2014;98:119–132. **DOI:** 10.1016/j.isprsjprs.2014.10.002

СЯНЬИ ВУ[1], АБЛАМЕЙКО С. В.[1,2]

МОДЕЛЬ ОБНАРУЖЕНИЯ ОБЪЕКТОВ НА ИЗОБРАЖЕНИЯХ ДИСТАНЦИОННОГО ЗОНДИРОВАНИЯ ЗЕМЛИ С ИСПОЛЬЗОВАНИЕМ ДИНАМИЧЕСКОГО РЕЦЕПТИВНОГО ПОЛЯ И SNAKE-СВЕРТКИ

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

Аннотация. *Для решения задачи обнаружения объектов на изображениях дистанционного зондирования Земли (ДЗЗ) в данной работе предлагается усовершенствованный метод на базе YOLOv11n. Предложена улучшенная архитектура YOLOv11n, интегрирующая модуль динамического рецептивного поля (RFAConv) и модуль адаптивного моделирования деформаций Snake (DySnakeConv). Этот подход улучшает процесс выявления низкоуровневых признаков и оптимизирует адаптивное выделение границ объектов, повышая точность обнаружения объектов. Эксперименты на наборе данных RSOD показали, что улучшенная модель достигает средней точности (mAP) 96.9 % при IoU = 0.50 (mAP50) и 65.5 % в диапазоне IoU 0.50–0.95 (mAP50-95). Результаты превосходят показатели YOLOv8n, YOLOv10n и других конкурентных моделей по ключевым метрикам (точности и полноте). Важно отметить, что модель сохраняет сопоставимую эффективность на наборе NWPU VHR-10. Предложенная модель является эффективным решением для обнаружения малых объектов и геометрически сложных целей на изображениях ДЗЗ высокого разрешения.*

Ключевые слова: *изображения дистанционного зондирования Земли, обнаружение объектов, YOLOv11, RFAConv, DySnakeConv*

By Сяньи
Белорусский государственный университет, Минск, 220030, Республика Беларусь.
Аспирант механико-математического факультета Белорусского государственного университета.

Xiangyi Wu
Belarusian State University, Minsk, 220030, Republic of Belarus.
Postgraduate student of the Faculty of Mechanics and Mathematics of the Belarusian State University.

E-mail: tigerv5872@gmail.com

Абламейко Сергей Владимирович
Объединенный институт проблем информатики Национальной академии наук Беларуси
Академик Национальной академии наук Беларуси, доктор технических наук, профессор. Лауреат Государственной премии Республики Беларусь.

Sergey Vladimirovich Ablameyko
The United Institute of Informatics Problems of the National Academy of Sciences of Belarus, Minsk, 220012, Republic of Belarus
Academician of the National Academy of Sciences of Belarus, Doctor of Science (Engineering), Professor. Laureate of the State Prize of the Republic of Belarus.

E-mail: ablameyko@bsu.by