scientific reports

Check for updates

OPEN

SED-YOLO based multi-scale attention for small object detection in remote sensing

Xiaotan Wei, Zhensong Li✉ & Yutong Wang

Object detection is crucial for remote sensing image processing, yet the detection of small objects remains highly challenging due to factors such as image noise and cluttered backgrounds. In response to this challenge, this paper proposes an improved network, named SED-YOLO, based on YOLOv5s. Firstly, we leverage Switchable Atrous Convolution (SAC) to replace the standard convolutions in the original C3 modules of the backbone network, thereby enhancing feature extraction capabilities and adaptability. Additionally, we introduce the Efficient Multi-Scale Attention(EMA) mechanism at the end of the backbone network to enable efficient multi-scale feature learning, which reduces computational costs while preserving crucial information. In the Neck section, an adaptive Concat method is designed to dynamically adjust the feature fusion strategy according to image content and object characteristics, strengthening the model's ability to handle diverse objects. Lastly, the three-scale feature detection head is expanded to four by adding a small object detection layer, and incorporating the Dynamic Head(DyHead) module. This enhances the detection head's expressive power by dynamically adjusting attention weights in feature maps. Experimental results demonstrate that this improved network achieves an mean Average Precision (mAP) of 71.6% on the DOTA dataset, surpassing the original YOLOv5s by 2.4%, effectively improving the accuracy of small object detection.

Keywords Remote sensing, Object detection, YOLO, Attention mechanism

Remote sensing imagery, as an efficient Earth observation technology, plays a key role in various fields such as disaster warning[1], intelligent transportation[2], aerospace[3], and intelligent surveillance[4]. Known for its high resolution and complex scenes[5], remote sensing images often contain a large number of small objects. These small objects typically occupy few pixels[6], exhibit significant scale variation, and are prone to interference from complex backgrounds and noise[7,8]. Traditional object detection methods often struggle to effectively extract object features in remote sensing images, and tend to have high computational complexity. In contrast, deep learning models, by constructing end-to-end training frameworks, can directly learn features and patterns from raw remote sensing imagery, reducing the need for manual feature extraction. As such, deep learning has become the primary method for the interpretation, analysis, and application of remote sensing images[9,10]. Although deep learning technology can efficiently extract key information from images, existing models still face significant challenges in small object detection[11]. For instance, the background of remote sensing images typically contains various land features, such as buildings, roads, trees, and rivers[12]. The challenge lies in accurately separating small objects from these complex backgrounds. Additionally, the objects in remote sensing images exhibit large scale differences, ranging from tiny ground objects to large structures such as buildings and vehicles. The key issue is how to maintain the integrity of small object information during multi-scale feature fusion and prevent accuracy loss. Therefore, designing an improved model that possesses efficient multi-scale perception capabilities and strong background noise suppression abilities to meet the specific needs of small object detection remains an important research direction.

Object detection models are generally categorized into two types: two-stage and single-stage models. Two-stage models, like R-CNN[13], Mask R-CNN[14], and Faster-RCNN[15] series, enhance accuracy through region proposal and precise detection stages, albeit at the cost of processing speed. Conversely, single-stage models, including YOLO series[16–21], SSD[22], and EfficientNet[23], directly predict categories and bounding boxes, offering faster and more efficient detection suitable for real-time applications. Although single-stage models possess an advantage in terms of detection speed, they often exhibit inferior performance compared to two-stage models when it comes to detecting small-scale objects or those situated within complex backgrounds. In order to mitigate this drawback, researchers have been persistently refining the precision and robustness of single-stage models.

Key Laboratory of the Ministry of Education for Optoelectronic Measurement Technology and Instrument, Beijing Information Science and Technology University, Beijing 100192, China. ✉email: lizhensong@bistu.edu.cn

For example, the incorporation of Feature Pyramid Networks (FPN)[24] serves to augment the model's capacity to perceive objects across various scales; the utilization of attention mechanisms[25] elevates the model's ability to concentrate on regions of interest within the object area; refinements to the anchor box strategy[26] enable more precise predictions regarding the dimensions of the object. Additionally, the employment of multi-scale feature fusion techniques bolsters the detection of small-scale objects[27]. The YOLO model, in particular, has integrated these sophisticated advancements, thereby continuously evolving to enhance its detection capabilities.

As deep learning progresses, YOLO series models continue to develop, offering robust technical support and innovative solutions for remote sensing image object detection. Xie et al.[28] introduced CSPPartial-YOLO, a lightweight model that improves detection accuracy and resource efficiency by incorporating partial mixed dilated convolution blocks and coordinate attention modules while reducing computational costs. Zhang et al.[29] proposed FFCA-YOLO, featuring enhanced local perception, multi-scale feature fusion, and global association capabilities through innovative feature enhancement, fusion, and spatial context awareness modules, effectively amplifying small object representation and suppressing background interference for efficient and accurate small object detection. Liu et al.[30] optimized YOLOv5 with YOLO-extract, integrating residual modules, coordinate attention mechanisms, removing underperforming feature layers and prediction heads, and incorporating new efficient feature extractors. Lin et al.[31] developed YOLO-DA, appending an attention module to the detector's end to prioritize efficient features and mitigate complex background interferences, alongside a lightweight decoupled detection head for superior localization and classification, thereby significantly boosting the performance of small object detection. Shi et al.[32] introduced LSKF-YOLO, leveraging an improved large spatial selection kernel mechanism and multi-scale feature alignment fusion to elevate power tower detection accuracy in complex satellite imagery, improving the detection accuracy of small objects in complex background high-resolution satellite remote sensing images. Lin et al.[33] presented GDRS-YOLO, integrating SPD convolution and an aggregate-distribute-based multi-scale feature aggregation network within YOLOv7's architecture, enhancing object geometry capture while minimizing information loss. Zhao et al.[34] proposed the YOLO-FSD remote sensing object detection algorithm, which introduces the Swin-CSP structure at each layer of the network to learn both local and global attributes, thereby enhancing the model's ability to discriminate objects in complex backgrounds with unclear boundaries. Additionally, a new DWC detection head is employed to reduce prediction bias caused by small and dense objects, while improving the localization and classification capabilities for small objects in complex backgrounds. Liu et al.[35] proposed a lightweight and efficient object detector based on YOLOv8n, which incorporates multi-scale dilated attention (MSDA) after the multi-scale feature fusion module to increase the model's focus on effective features in complex backgrounds. Furthermore, they adopted Shape-IoU as the bounding box regression loss to improve the model's localization accuracy.

Despite these advancements, we still face challenges such as insufficient utilization of small object information, inadequate feature extraction, and weak robustness against complex backgrounds and noise. For instance, references[28,30,31] enhance object feature representation and suppress background interference by incorporating coordinate attention modules. However, due to the high similarity between objects and backgrounds in remote sensing images, the attention mechanism may struggle to accurately distinguish small objects from background information in complex environments. Furthermore, the introduction of attention modules adds additional computational overhead, which can lead to a decrease in model inference speed. References[32–35] improve object detection performance through multi-scale feature fusion. While these methods effectively enhance multi-scale feature representation, the fusion process often introduces redundant features[36]. Moreover, during multi-scale fusion, misalignment between shallow and deep semantic features may occur, potentially leading to inconsistent feature representation and negatively impacting the precise detection of small objects.

Inspired by these limitations and prior literature, this paper introduces an innovative YOLOv5s-based detection method SED-YOLO, which notably improves small object detection performance, enhancing both detection accuracy and speed. Our key contributions are:

- Backbone Network Enhancements: Replacing standard convolutions in C3 with Switchable Atrous Convolution (SAConv)[37], bolstering feature capture across diverse scales through a combination of multi-sized kernels and attention mechanisms. Additionally, appending Efficient Multi-Scale Attention (EMA)[38] module at the backbone's end significantly enhances small object detection, robustness to complex backgrounds and noise, and generalization across scale objects.
- Adaptive Concat in Neck Section: Designing an adaptive Concat method that dynamically adjusts feature fusion based on input feature maps' semantic and spatial resolutions, preserving high-dimensional information, augmenting small object detection, and minimizing computational resources.
- Dynamic Head (DyHead)[39] Module and Expanded Detection Heads: Introducing DyHead and an additional small object detection head, enhancing the model's dynamic feature capture and representation capabilities, significantly improving detection head performance.

Ultimately, an experimental study on the detection of small objects in remote sensing images was conducted using the DOTA dataset[40], and the proposed model was compared with several excellent YOLO models. The results demonstrated that the model exhibits superior performance in terms of accuracy, size, and speed. The remainder of this paper is organized as follows: the second section describes the structure of the improved model and specific improvement modules of the model, the third section presents the relevant experimental results and analysis, and finally, the main findings are summarized and research conclusions are presented.

Improved algorithm

Inspired by YOLOv5s, we propose an improved neural network, named SED-YOLO, aimed at further enhancing the detection capabilities for small objects and complex backgrounds in object detection tasks.

In the backbone network part, the standard convolution in the C3 module is replaced with SAConv and renamed C3_SAC. At the same time, the EMA module is incorporated at the end of the network, which improves the feature extraction capability of the backbone network and enhances the model's performance in small objects and complex backgrounds; In the Neck part of the model, we designed an adaptive Concat method that dynamically adjusts the feature fusion strategy according to the multi-scale characteristics of the input features, effectively integrating semantic and spatial information from different levels; In addition, in the Head part, by adding a dedicated small object detection layer and introducing the DyHead module, the expression ability of the detection head is significantly improved. This series of innovative network designs and optimization strategies have achieved significant performance improvements on standard datasets.

To further elaborate on how the aforementioned improvements are embodied in the actual network structure, we have designed an improved model architecture diagram, as shown in Fig. 1. Next, we will divide the discussion into three parts: Backbone Network Enhancements, Neck Improvements, and Detection Head Improvements, to delve into the key components of our proposed improved neural network and their roles.

Backbone network enhancements

In object detection models, the ability of the backbone network to extract information plays a crucial role in the overall model's detection accuracy. To address the challenge of capturing numerous small objects and their detailed information in remote sensing images, SAConv is used to replace the standard C3 convolution in the shallow layers of the backbone network, while the EMA attention mechanism is integrated at the end of the backbone network. This approach significantly enhances the detection capability for small objects and complex scenes, demonstrating stronger generalization and robustness of the model.

SAConv

The SAConv architecture leverages the advantages of Atrous Convolution, enabling it to capture object information at various scales by expanding the receptive field, while preserving image resolution. This characteristic allows SAConv to excel in handling multi-scale objects in remote sensing images. Additionally, the dynamic feature fusion strategy of SAConv introduces an attention mechanism (with weight parameters S and 1 –S), adaptively weighting features from dilated convolutions with different dilation rates. This enables the model to flexibly

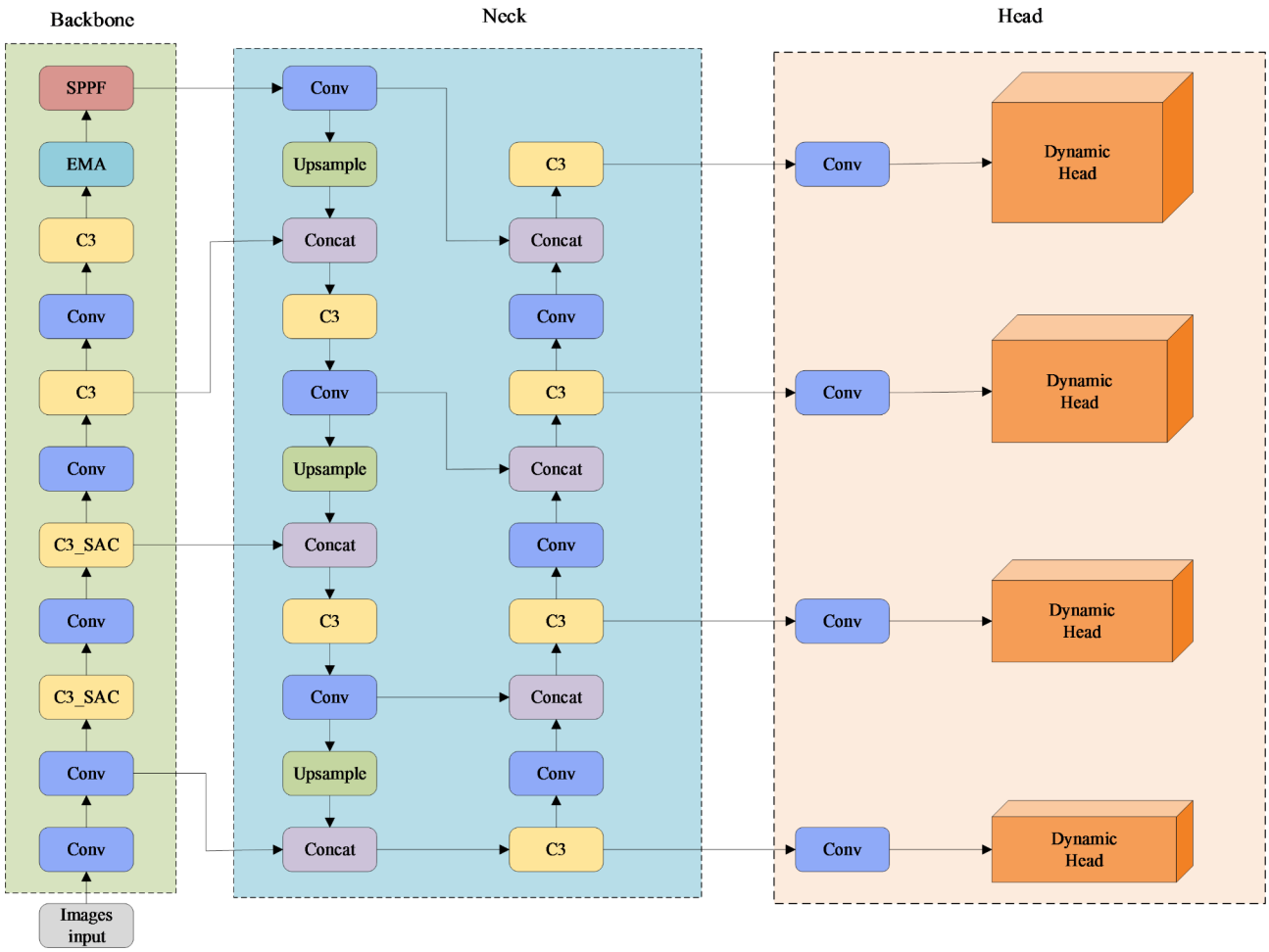

Fig. 1. Improved Yolov5s architecture.

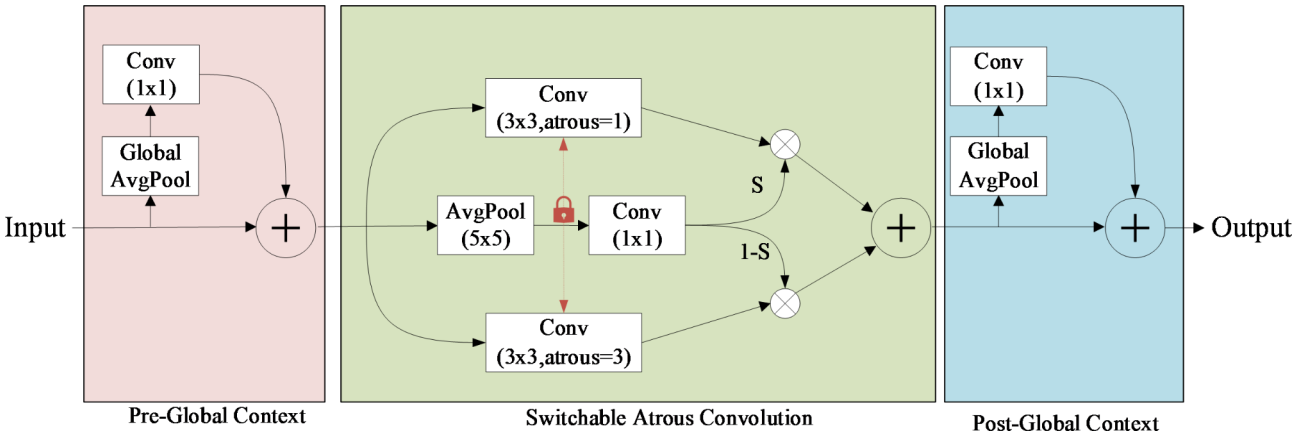

Fig. 2. SAC module structure.

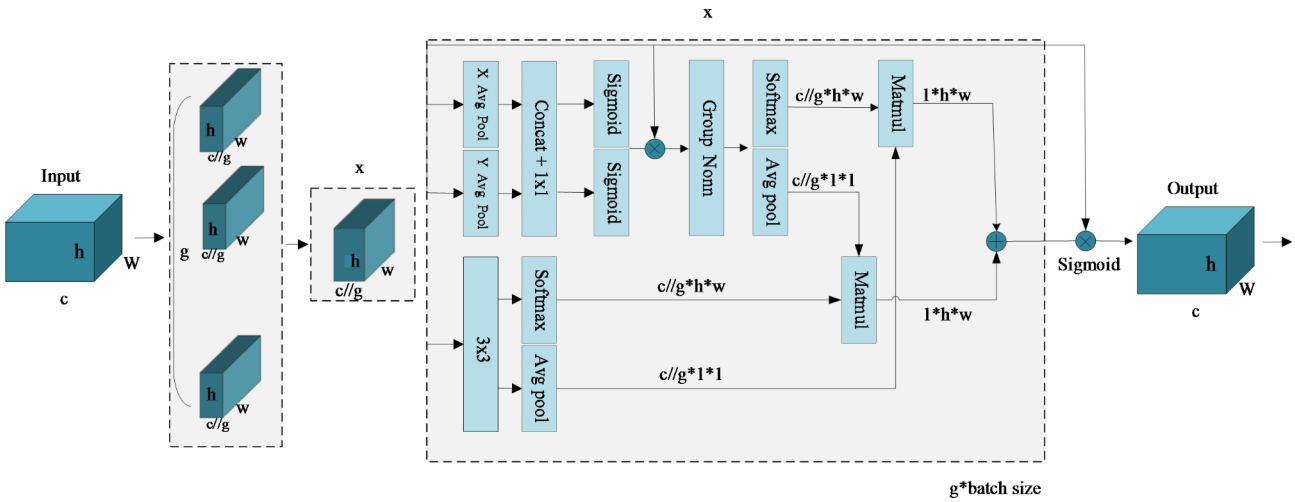

Fig. 3. The structure of EMA.

adjust its feature extraction strategy based on the characteristics of the remote sensing image scene, allowing it to effectively respond to the diversity of objects in remote sensing images. Through this strategy, the model can automatically adjust its feature extraction method according to the image scene, effectively addressing the challenges of small object detection in complex backgrounds and reducing background noise interference. Figure 2 illustrates the structure of the SAC module.

In this figure, the SAC module consists of three main components. The first part is the Pre-Global Context, which extracts global contextual information from the input features using a 1×1 convolution and global average pooling, and then adds it to the original input to enhance the global perception capability of shallow features. The second part is the Switchable Atrous Convolution, where two 3×3 dilated convolutions with different dilation rates (atrous rate = 1 and 5) are employed to capture local and global features at different scales. Finally, the Post-Global Context processes the features again using a 1×1 convolution and global average pooling, further enhancing the fusion of local and global features, and combines this with the output from the previous part to obtain the final result.

EMA module

The addition of the EMA attention mechanism at the end of the YOLOv5s backbone network aims to enhance the model's attention to feature information across different scales, thereby improving its ability to detect multi-scale objects. The core advantage of the EMA mechanism lies in its ability to perform attention operations in parallel across multiple scales, accurately capturing key information in the image. Figure 3 illustrates the structure of the EMA mechanism, detailing its specific implementation process. First, the input feature map $C \times H \times W$ is divided into multiple groups (e.g., groups), with each group handling a portion of the channels, reducing computational complexity. Next, the input feature map is downsampled at different scales (e.g., scales of 1/2 and 1/4) through pooling operations, generating feature maps with different resolutions that capture both global and local information. These multi-scale feature maps are then processed in parallel, and the multi-

scale information is fused to generate an attention weight map. The attention weights are normalized using the Sigmoid activation function, and the final fused feature map outis obtained. This feature map retains the same spatial dimensions H×W as the input, but with enhanced multi-scale dynamic perception capabilities through the optimization provided by the attention mechanism.

Given the limited proportion and highly variable scale of small objects in remote sensing images, the EMA mechanism, through the combination of these strategies, effectively captures the detailed features of small objects while retaining the global contextual information of larger objects. By generating attention weights, it adaptively weights features at different scales, highlighting object regions and suppressing background noise interference. This significantly enhances the saliency of objects in complex backgrounds, thereby reducing both false negatives and false positives in object detection.

Neck improvements

In the Neck section, due to the insufficient proportion of small object information, it is challenging to fully utilize global context when fusing multi-scale features, leading to reduced accuracy in predicting object position, shape, and category. To address this issue, we designed an adaptive Concat module that dynamically adjusts the fusion strategy based on the size and content of the input feature maps, enhancing the model's ability to capture details and contextual information. Through the network's adaptive adjustment and efficient feature fusion, we achieved effective integration of deep and shallow features extracted from the backbone, optimizing the detection performance for small objects and ensuring that the proportion of small object information is maximized.

Figure 4 illustrates the adaptive Concat structure with two and three input branches. As shown, the module first selects the corresponding weight parameters based on the number of input feature maps. The weighted feature maps are then concatenated along a predefined dimension using the Concat function, facilitating feature integration. Through this adaptive feature fusion strategy, the model can better capture objects of various sizes, thus maintaining high detection performance in diverse and complex scenarios. Additionally, the adaptive Concat module efficiently allocates computational resources by selectively integrating key feature maps, rather than performing a uniform fusion operation across all input feature maps, thereby reducing computational overhead.

Detection head improvements

Adding an additional detection layer

The traditional YOLOv5s model includes three detection heads, each corresponding to feature maps of different scales to handle large, medium, and small objects. However, the original model still has certain limitations in detecting small objects. To address this issue, we have added an extra small object detection layer to the Head section of YOLOv5s, resulting in a total of four detection heads. By introducing this new detection head, the module can better utilize multi-scale information and detailed features within the feature maps, thereby improving its performance in detecting small objects.

Specifically, the newly added small object detection layer enhances the model's ability to perceive small objects through more fine-grained feature extraction and processing. The new detection head works synergistically with the existing detection heads, fully leveraging the multi-scale information in the feature maps, making the model more efficient in detecting objects of varying sizes. This design not only increases the accuracy of small object detection but also enhances the model's overall detection performance in complex scenarios, providing a more reliable solution for multi-scale object detection tasks.

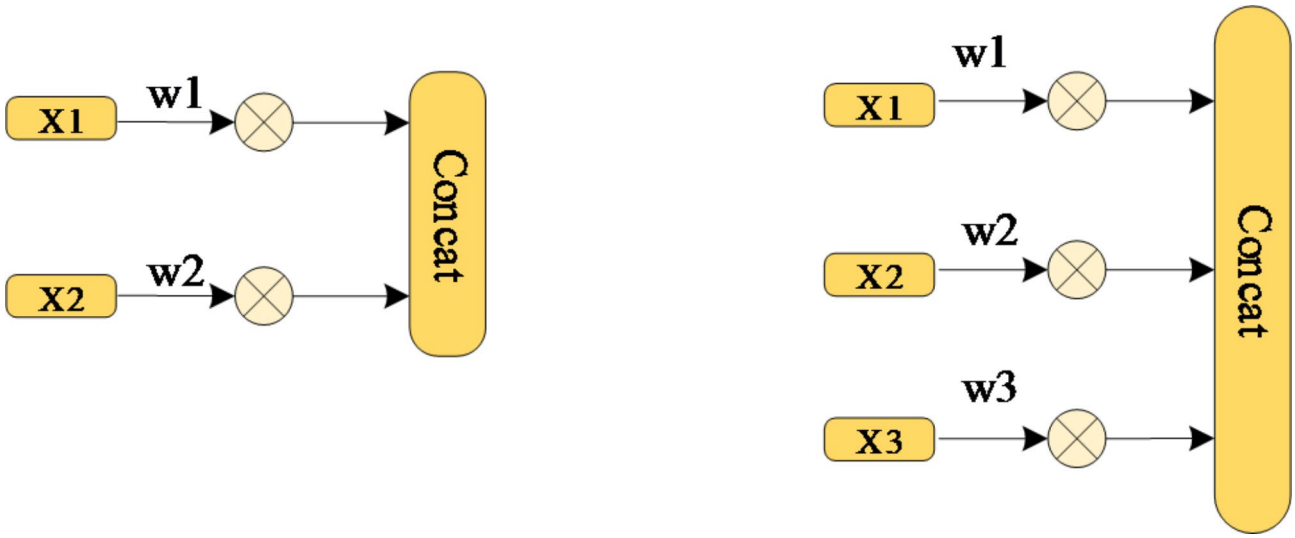

Fig. 4. Adaptive Concat structure.

DyHead module

DyHead module is an innovative technique in the field of object detection designed to enhance model performance. Its specific structure and the relationships between its components are illustrated in Fig. 5. To alleviate the burden of high-dimensional computations, the DyHead module decomposes the attention mechanism into three sequential operations. The first mechanism is the scale-aware attention mechanism, which effectively captures the local features of different object categories by adaptively weighting features at various scales. The second mechanism is the spatial-aware attention mechanism, which dynamically assigns different weights to different spatial locations, emphasizing small object regions while suppressing redundant background and noise interference. Lastly, the task-aware attention mechanism dynamically adjusts the feature representation based on different detection tasks, enhancing the multi-task performance of object detection in remote sensing images and ensuring more accurate localization of small objects.

The DyHead module takes multi-scale feature maps from theFPN as input and reorganizes and processes these feature maps across different dimensions (S: Scale, L: Location, C: Channel). By integrating these three attention mechanisms, the object detection head is able to more accurately and robustly handle multi-scale object detection tasks. Compared to traditional single attention mechanisms, DyHead demonstrates significant performance improvements, particularly in detecting small objects in complex backgrounds. It effectively reduces false positives and false negatives, providing a more reliable solution for object detection tasks.

Overall, a significant improvement in detection performance is achieved in the Head section by adding a small object detection layer and introducing the DyHead module. This improvement not only improves the recognition of small objects through multi-scale feature fusion, but also further improves the recognition of small objects and enhances the generalisation ability of the model by using the dynamic feature fusion mechanism of the DyHead module. This enables YOLOv5 to significantly improve its accuracy in the small object detection task while maintaining a relatively high processing speed.

Experiment results and analysis

Dataset

The DOTAV1.0 dataset is a large-scale remote sensing image dataset specifically designed to evaluate the performance of object detection algorithms in aerial imagery. This dataset contains 2806 aerial images, with pixel dimensions ranging from 800×800 to 4000×4000, covering a wide range of geographical scenes and diverse object categories. It includes a total of 188,282 object instances, which are meticulously categorized into 15 common classes, such as airplanes, ships, tanks, baseball fields, tennis courts, etc. Each instance is annotated with precise quadrilateral bounding boxes, ensuring the accuracy and usability of the data. To enhance model training effectiveness, the original images are preprocessed through image segmentation and padding, expanding the dataset from 2806 images of varying resolutions to 21,046 images, with 15,749 used for training and 5297 for testing.

Experimental environment

The experiments were conducted on a Windows 10 operating system with the hardware configuration including an RTX 4090 (24GB) GPU. The software environment used for the experiments included Python 3.8, PyTorch 1.11.0, and Cuda 11.3. Under identical hyperparameters, training, validation, and testing were carried out with a batch size set to 16 and a learning rate of 0.01. The training utilized the Stochastic Gradient Descent (SGD) optimizer, and the image resolution in the dataset was 640×640. The number of epochs was set to 350, and the Complete Intersection over Union (CIOU) loss function was chosen.

Evaluation metrics

To objectively evaluate the model's detection performance, we used mean Average Precision (mAP) as the core evaluation metric. mAP is a commonly used metric to assess the overall performance of object detection models, calculated as the average of the Average Precision (AP) across all classes. AP is typically computed as the area under the Precision-Recall curve, which is generated by varying the confidence threshold. Precision reflects the proportion of correctly identified objects by the model, while recall indicates the model's ability to capture all objects. The formulas for calculating these evaluation metrics are as follows:

$$P = \frac{T_P}{T_P + F_P} \times 100\% \tag{1}$$

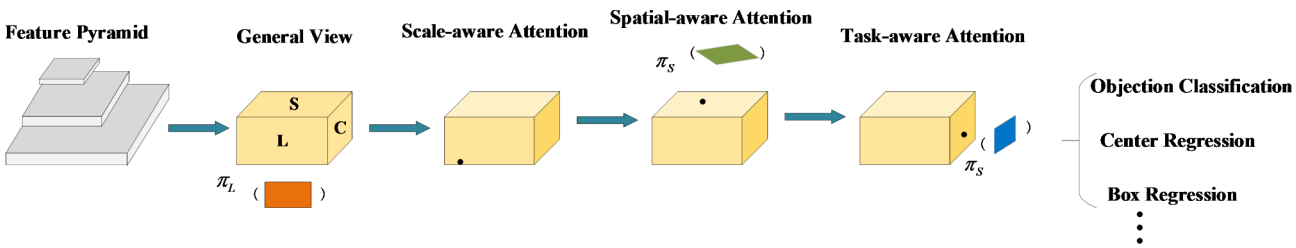

Fig. 5. DyHead overall structure.

$$R = \frac{T_P}{T_P + F_N} \times 100\% \tag{2}$$

Among these, is the number of samples that the model correctly predicts as positive when they are actually positive. is the number of samples that the model incorrectly predicts as positive when they are actually negative. is the number of samples that the model incorrectly predicts as negative when they are actually positive.The formulas for calculating AP and mAP are as follows:

$$AP = \int_0^1 P(R)dR \tag{3}$$

$$mAP = \frac{1}{n}\sum_{i=1}^{n} AP_i \tag{4}$$

In addition to pursuing model detection performance, we also place significant emphasis on the model's efficiency. Therefore, we select GFLOPs (Giga Floating-Point Operations per Second) and Params (parameter count) to evaluate the computational resource requirements and model scale. Inference Time (IT) serves as a key metric for assessing the model's real-time performance. By considering these metrics comprehensively, we can achieve a balanced evaluation of the model's performance and efficiency.

Ablation studies

To evaluate the impact of the proposed improvements on the performance of the YOLOv5s model, we designed a series of ablation studies. By individually removing or adding each improvement component—such as the SAConv, EMA module, adaptive Concat, and the four detection heads with the DyHead module—we can analyze the contribution of each component to the model's detection performance. The table below presents the specific results of the ablation experiments.

As shown in Table 1, our optimization strategies effectively enhance performance. In the baseline experiment, the original YOLOv5s model was used for detection, with a parameter size of 7.05 M, a computational load of 15.9 GFLOPs, and an inference time of 2.1 ms. Compared to other models, it has a lower computational cost, but its mAP is also the lowest at 69.2%.

After improving the C3_SAC module in the backbone network, the model's mAP increased by 0.7%. With the addition of the EMA module, as shown in Fig. 6, we observed an enhancement in the model's focus on objects and detection accuracy, with a 0.9% increase in mAP. The increase in model parameters was relatively small compared to other modules. Following the introduction of the adaptive Concat module, the model's mAP improved by 0.6% compared to the baseline, while the computational load was reduced, and the inference time remained largely unchanged.

Although the addition of four detection heads with the DyHead module increased the computational load, it significantly enhanced the model's accuracy, bringing the mAP up to 71.0%, a 1.8% improvement over YOLOv5s. The parameters and computational load were 7.96 M and 36.8 GFLOPs, respectively, with an inference time of 4.2 ms. This further strengthened the model's ability to extract information and achieve superior detection performance. These results indicate that our optimization strategies not only enhance the model's detection performance but also achieve an effective balance between performance and complexity.

Visualization results

As shown in Fig. 7, to more clearly demonstrate the effectiveness of the improved modules, we used visualized detection results. We selected four different scenarios from the DOTA dataset for comparison. The first column of sub-images shows the ground truth labels, the second column displays the detection results from the YOLOv5s model, and the third column shows the detection results from our model.

In the first row, which focuses on the detection of densely packed vehicles on roads, the comparison images show that the YOLOv5s model failed to capture the small objects within the blue circles, whereas our model successfully detected the small objects that YOLOv5s missed. The second row involves the detection of small objects in an airport tarmac scenario. It can be observed that the YOLOv5s model failed to detect certain markers around the airplanes, while our model successfully identified these small objects missed by YOLOv5s. The third row of images also shows detection results from an airport tarmac scenario. In this case, the YOLOv5s model exhibited false detections, incorrectly marking some non-existent objects around the airplanes. Our method not

C3_SAC	EMA	Adaptive concat	Four-head DyHead	mAP_{50} (%)	IT (ms)	Params (M)	GFLOPs
–	–	–	–	69.2	2.1	7.05	15.9
✓	–	–	–	69.9	2.2	7.15	14.8
–	✓	–	–	70.1	2.3	7.09	16.4
–	–	✓	–	69.8	2.2	7.06	15.8
–	–	–	✓	71.0	4.2	7.96	36.8
✓	✓	✓	✓	71.6	4.0	8.09	35.3

Table 1. Ablation study results.

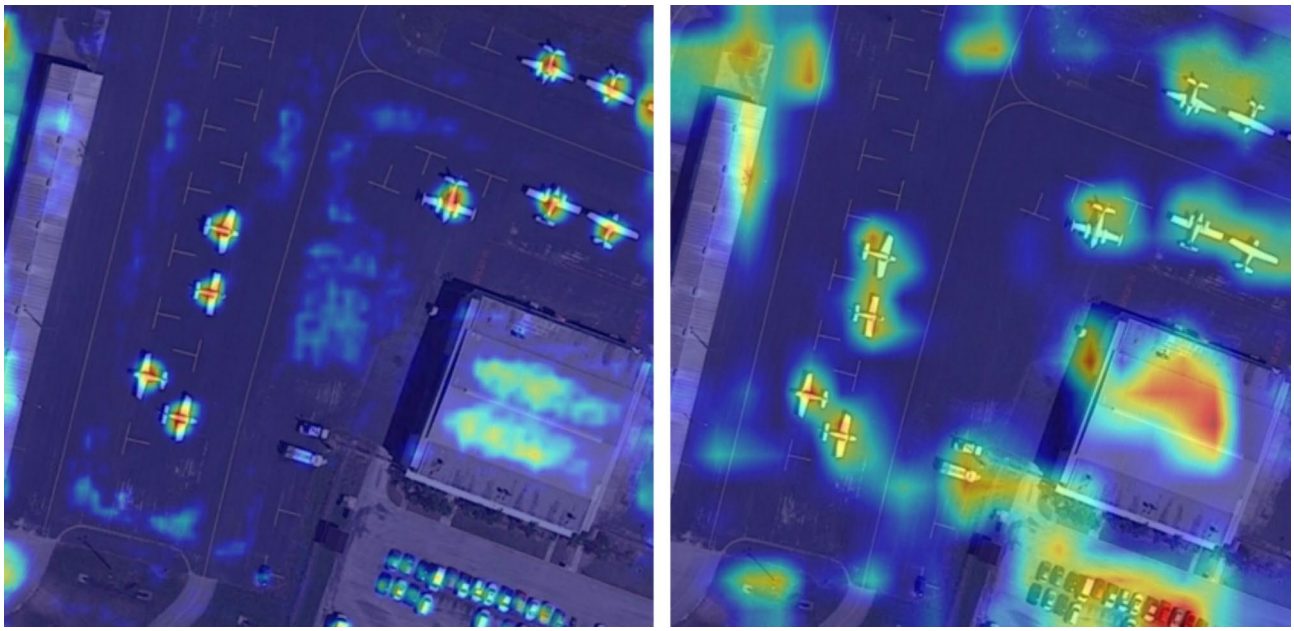

Fig. 6. EMA module contrast heat map.

only accurately identified the small objects on the tarmac but also effectively avoided false detections. In the final coastal scene, the YOLOv5s model failed to detect the vehicles on the shore within the blue circles, particularly in the complex backgrounds along the shore or water's edge.

In summary, after adding the adaptive Concat module and the four detection heads with the DyHead module, our model achieved notable improvements. It reduced overfitting and decreased both false positives and missed detections. This enhanced precision and robustness make our model more effective in handling complex image recognition tasks.

Comparative experiments

In order to further comprehensively evaluate the performance of the improved YOLOv5s model, we designed a series of comparison experiments. By comparing our model with several mainstream object detection models, such as YOLOv3-tiny, YOLOv5s baseline model, RetinaNet, and other improved models, we aim to verify the effectiveness and superiority of the improved model in practical applications. The experiments use the same dataset and evaluation metrics, and compare in detail the performance of different models in terms of computational resources, number of parameters and mAP. Table 2 lists the specific results of these experiments on the DOTA dataset, with the best results in each category highlighted in bold.

The results indicate that YOLOv3-tiny, as a lightweight model, has lower GFLOPs and parameter count but correspondingly lower mAP. In comparison with references[37–40], SED-YOLO achieves a higher mAP while maintaining a relatively low parameter count. In terms of individual class detection accuracy, SED-YOLO performs at its best for categories such as Storage-tank, Baseball-diamond, Basketball-diamond, Ground-track field, and Helicopter, with over half of the categories reaching optimal or near-optimal levels. Overall, the analysis shows that SED-YOLO demonstrates exceptional performance, surpassing most other object detection models.

Conclusions

This paper proposes SED-YOLO, an improved YOLOv5s model designed for small object detection in remote sensing images. By introducing SAConv into the backbone network and integrating an EMA mechanism at the end, the model's robustness against complex backgrounds and its ability to detect small objects have been significantly enhanced. Additionally, an adaptive Concat strategy has been designed in the Neck section to dynamically adjust the feature fusion method, further improving the efficiency of multi-scale feature utilization. Moreover, by adding detection heads and combining them with the DyHead module, the model demonstrates superior performance in multi-scale object detection tasks. Experimental results on the DOTA dataset validate the effectiveness of these improvements, with SED-YOLO achieving a remarkable 2.4% increase in mAP compared to the original YOLOv5s. This advancement not only improves detection accuracy but also maintains computational efficiency, making it highly suitable for real-time applications in remote sensing. The structure of SED-YOLO provides a solid foundation for future research, it is particularly significant for the further development of small object detection, especially in detecting objects like storage tanks, baseball diamonds, basketball courts, and so on.

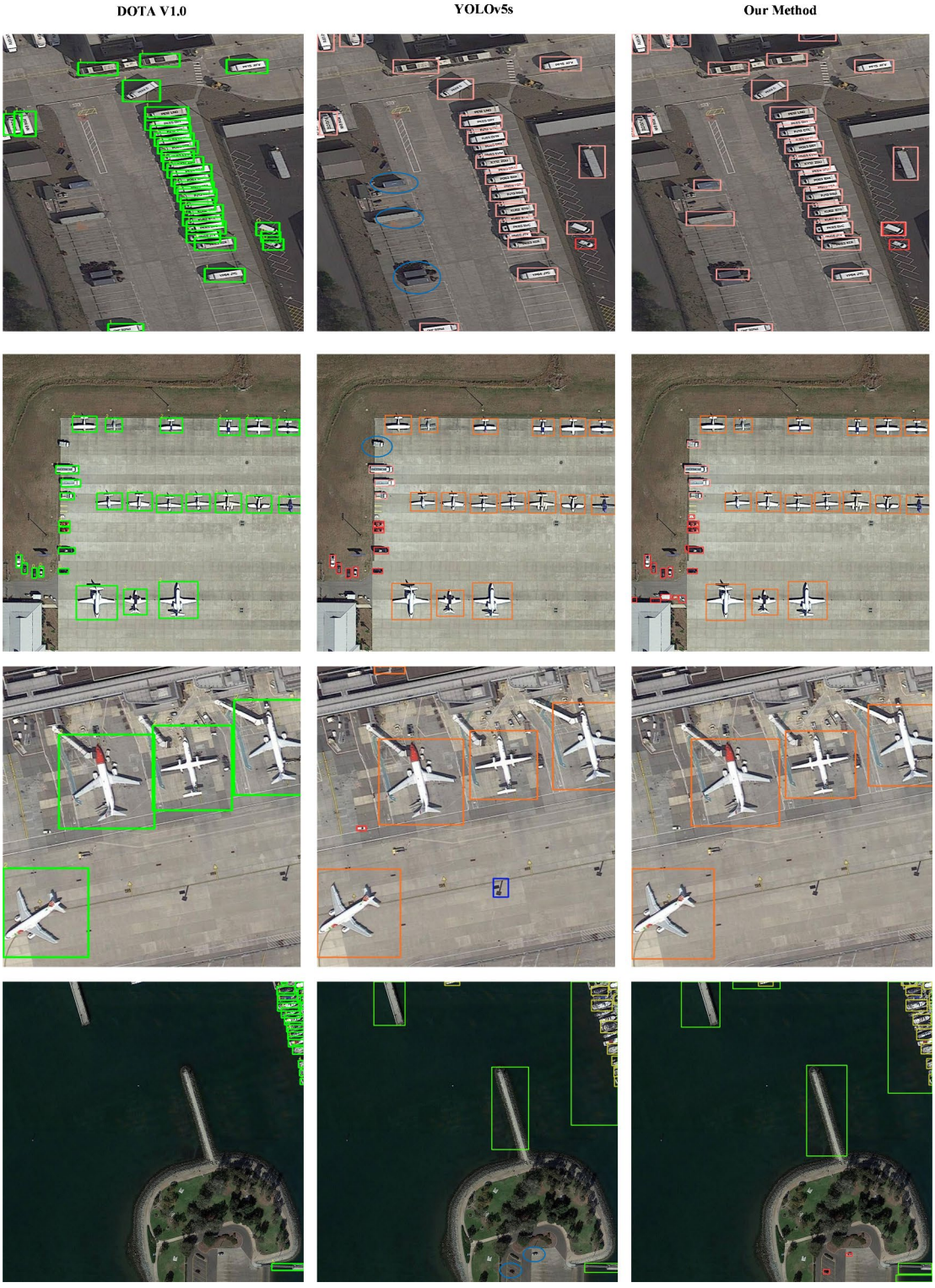

Fig. 7. Visualization results.

Model	YOLOv3- tiny	YOLOv5s	RetinaNet	ARSD[41]	MDCNet[42]	CoF-Net[43]	LO-Det[44]	Ours
Plane	42.1	67.5	74.6	86.9	92.7	89.7	89.2	70.0
Ship	68.3	85.4	52.1	79.0	88.5	82.1	84.2	86.4
Storage-tank	86.9	92.0	37.2	79.0	60.2	83.9	81.3	**92.8**
Baseball-diamond	46.6	70.4	33.6	72.2	80.6	55.5	66.1	**80.9**
Tennis-court	66.9	86.7	61.2	90.7	**92.9**	89.6	90.7	89.3
Basketball-diamond	75.6	82.7	51.6	72.2	80.6	72.9	75.1	**83.9**
Ground-track field	43.6	58.0	34.1	56.9	56.6	69.3	56.0	**58.2**
Harbor	48.8	51.8	32.1	72.2	**83.0**	54.5	59.9	44.7
Brige	90.7	92.8	21.2	46.4	48.5	41.6	31.3	**93.3**
Small vehicle	51.4	60.5	38.4	**80.6**	65.5	65.1	65.1	62.2
Large vehicle	63.3	74.9	29.7	67.0	**86.5**	77.2	71.0	77.1
Helicopter	43.3	59.3	13.3	50.6	35.1	59.0	48.4	**59.7**
Roundabout	45.2	60.9	36.8	61.3	59.3	**66.1**	59.3	58.3
Soccer-ball field	32.4	47.0	29.4	**79.0**	59.8	72.8	44.7	50.9
Swimming pool	42.3	54.9	38.2	**72.4**	68.8	58.8	54.9	66.4
mAP_{50}(%)	56.5	69.2	38.9	68.3	69.2	69.3	66.2	**71.6**
IT(ms)	2.2	2.3	28.7	23.2	—	47.6	16.7	4.0
Params(M)	8.7	7.0	36.4	13.1	—	37.6	26.9	8.0
GFLOPs	12.9	15.9	213.6	68.0	—	—	—	35.3

Table 2. Comparison of detection results with other methods.

Data availability

The data generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Received: 2 September 2024; Accepted: 16 January 2025
Published online: 24 January 2025

References

1. Zhang, W. et al. LS-YOLO: a novel model for detecting multiscale landslides with remote sensing images. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **17**, 4952–4965 (2024).
2. Patil, P. Applications of deep learning in traffic management: a review. *Int. J. Bus. Intell. Big Data Anal.* **5**(1), 16–23 (2022).
3. Xu, Y., Zhu, M., Xin, P., Li, S., Qi, M. & Ma, S. Rapid airplane detection in remote sensing images based on multilayer feature fusion in fully convolutional neural networks. *Sensors* **18**, 7 (2018).
4. Wang, E. K., Wang, F., Kumari, S., Yeh, J.-H. & Chen, C.-M. Intelligent monitor for typhoon in IOT system of smart city. *J. Supercomput.* **77**(3), 3024–3043 (2021).
5. Han, H., Zhu, F., Zhu, B. & Wu, H. Target detection of remote sensing image based on an improved YOLOv5. *IEEE Geosci. Remote Sens. Lett.* **20**, 1–5 (2023).
6. Sun, L. et al. CRNet: channel-enhanced remodeling-based network for salient object detection in optical remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **61**, 1–14 (2023).
7. Ruan, H., Qian, W., Zheng, Z. & Peng, Y. A decoupled semantic-detail learning network for remote sensing object detection in complex backgrounds. *Electronics* **12**(14), 3201 (2023).
8. Ran, Q., Wang, Q., Zhao, B., Wu, Y., Pu, S. & Li, Z. Lightweight oriented object detection using multiscale context and enhanced channel attention in remote sensing images. *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.* **14**, 5786–5795 (2021).
9. Wang, Y., Li, S., Teng, F., Lin, Y., Wang, M. & Cai, H. Improved mask R-CNN for rural building roof type recognition from UAV high-resolution images: a case study in Hunan province, China. *Remote Sens.* **14**(2), 265 (2022).
10. Wang, Y., Li, S., Lin, Y. & Wang, M. Lightweight deep neural network method for water body extraction from high-resolution remote sensing images with multisensors. *Sensors* **21**(21), 7397 (2021).
11. Khan, S. D., Alarabi, L. & Basalamah, S. A unified deep learning framework of multi-scale detectors for geo-spatial object detection in high-resolution satellite images. *Arab J. Sci. Eng.* **47**, 9489–9504 (2022).
12. Yan, P. et al. Clustered remote sensing target distribution detection aided by density-based spatial analysis. *Int. J. Appl. Earth Obs. Geoinformation* **132,** 104019 (2024).
13. Girshick, R., Donahue, J., Darrell, T. & Malik, J. Richfeaturehierarchies for accurate object detection and semantic segmentation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit* 580–587 (2014).
14. K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in Proc. IEEE Int. Conf. Comput. Vis, pp. 2961–2969(2017).
15. Ren, S., He, K., Girshick, R. & Sun, J. Faster R-CNN: towardsreal-time object detection with region proposal networks. In *Proc. Adv. Neural Inf. Process. Syst* (2015).
16. Redmon, J., Divvala, S., Girshick, R. & Farhadi, A. You only look once: unified, real-time object detection. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit* 779–788 (2016).
17. Redmon, J. & Farhadi, A. YOLOv3: an incremental improvement. arXiv:1804.02767 (2018).
18. Bochkovskiy, A., Wang, C.-Y. & Liao, H.-Y. M. YOLOv4: optimal speed and accuracy of object detection (2020).
19. Ge, Z., Liu, S., Wang, F., Li, Z. & Sun, J. YOLOX: exceeding YOLO series in 2021. arXiv:2107.08430v2 (2021).
20. Li, C., Li, L., Geng, Y., Jiang, H., Cheng, M., Zhang, B. et al. YOLOv6 v3.0: a full-scale reloading. arXiv:2301.05586 (2023).
21. Aboah, A., Wang, B., Bagci, U. & Adu-Gyamfi, Y. Real-time multi-class helmet violation detection using few-shot data sampling technique and YOLOv8. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.Workshops (CVPRW)* 5349–5357 (2023).
22. Liu, W. et al. SSD: single shot multibox detector. In *Proc. Eur. Conf. Comput. Vis* 21–37 (2016).

23. Tan, M., Pang, R. & Le, Q. V. EfficientDet: scalable and efficient object detection. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit* 10778–10787 (2020).
24. Lin, T. Y., Dollár, P., Girshick, R. B., He, K., Hariharan, B. & Belongie, S. J. Feature pyramid networks for object detection. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)* 2117–2125 (2017).
25. Hu, J., Shen, L. & Sun, G. Squeeze-and-excitation networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **42**(8), 2132–2148 (2020).
26. Liu, W. et al. Anchor box: a fast institution for object detection and segmentation. In *Proc. European Conf. Comput. Vis. (ECCV)* (2018).
27. Wang, Z., Wang, C., Li, X., Xia, C. & Xu, J. MLP-Net: multilayer perceptron fusion network for infrared small target detection. *IEEE Trans. Geosci. Remote Sens.* **63**, 1–13 (2025).
28. Xie, S., Zhou, M., Wang, C. & Huang, S. CSPPartial-YOLO: a lightweight YOLO-based method for typical objects detection in remote sensing images. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **17**, 388–399 (2024).
29. Zhang, Y. et al. FFCA-YOLO for small object detection in remote sensing images. *IEEE Trans. Geosci. Remote Sens. 62*, 1–15 (2024).
30. Liu, Z., Gao, Y., Du, Q., Chen, M. & Lv, W. YOLO-extract: improved YOLOv5 for aircraft object detection in remote sensing images. *IEEE Access* **11**, 1742–1751 (2023).
31. Lin, J., Zhao, Y., Wang, S. & Tang, Y. YOLO-DA: an efficient YOLO-based detector for remote sensing object detection. *IEEE Geosci. Remote Sens. Lett.* **20**, 1–5 (2023).
32. Shi, C. et al. LSKF-YOLO: large selective kernel feature fusion network for power tower detection in high-resolution satellite remote sensing images. *IEEE Trans. Geosci. Remote Sens.* 62, 1–16 (2024).
33. Lin, Y., Li, J., Shen, S., Wang, H. & Zhou, H. GDRS-YOLO: more efficient multiscale features fusion object detector for remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **21**, 1–5 (2024).
34. Zhao, H., Chu, K., Zhang, J. & Feng, C. YOLO-FSD: an improved target detection algorithm on remote-sensing images. *IEEE Sens. J.* **23**(24), 30751–30764 (2023).
35. Liu, F., Hu, W. & Hu, H. YOLO-RMS: a lightweight and efficient detector for object detection in remote sensing. *IEEE Geosci. Remote Sens. Lett.* **21**, 1–5 (2024).
36. Meng, S. et al. A robust infrared small target detection method jointing multiple information and noise prediction: algorithm and benchmark. *IEEE Trans. Geosci. Remote Sens.* **61**, 1–17 (2023).
37. Qiao, S., Chen, L. -C. & Yuille, A. DetectoRS: detecting objects with recursive feature pyramid and switchable atrous convolution. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA* (2021).
38. Ouyang, D. et al. Efficient multi-scale attention module with cross-spatial learning. In *ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece* 1–5 (2023).
39. Dai, X. et al. Dynamic Head: unifying object detection heads with attentions. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA* 7369–7378 (2021).
40. Xia, G. -S. et al. DOTA: a large-scale dataset for object detection in aerial images In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA* 3974–3983 (2018).
41. Yang, Y. et al. Adaptive knowledge distillation for lightweight remote sensing object detectors optimizing. *IEEE Trans. Geosci. Remote Sens.* **60**, 1–15 (2022).
42. S. Duan et al. MDCNet: a multiplatform distributed collaborative network for object detection in remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **62**, 1–15 (2024).
43. Zhang, C., Lam, K.-M. & Wang, Q. CoF-Net: a progressive coarse-to-fine framework for object detection in remote-sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **61**, 1–17 (2023).
44. Huang, Z. et al. LO-Det: lightweight oriented object detection in remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **60**, 1–15 (2022).

Acknowledgements

This work is partly supported by Joint Funds of the National Natural Science Foundation of China, (Ye Qisun Science Foundation) under Grant U2341223, and the National Natural Science Foundation of China under Grant 62074017.

Author contributions

W.X. was responsible for the overall content structure, experimental design and analysis of the paper. W.Y. was responsible for the overall layout of the paper, referencing, and embellishment. L.Z.was responsible for the overall direction and supervision of the paper. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Z.L.