

RESEARCH

Open Access



Anchor-free object detection in remote sensing images using a variable receptive field network

Shenshen Fu¹, Yifan He^{2*}, Xiaofeng Du^{1*} and Yi Zhu¹

*Correspondence:
heyifan@reconova.com;
xfdu@xmut.edu.cn

¹ Faculty of Computer
Science School of Computer
and Information Engineering,
Xiamen University of Technology,
No. 600, Polytechnic Road,
Xiamen 361024, China

² Institute of Intelligence Science
and Engineering, Shenzhen
Polytechnic, No. 7098, Liuxian
Avenue, Shenzhen 518055, China

Abstract

Object detection is one of the essential tasks in computer vision, with most detection methods relying on a limited number of sizes for anchor boxes. However, the boundaries of particular composite objects, such as ports, highways, and golf courses, are ambiguous in remote sensing images, and therefore, it is challenging for the anchor-based method to accommodate the substantial size variation of the objects. In addition, the dense placement of anchor boxes imbalances the positive and negative samples, which affects the end-to-end architecture of deep learning methods. Hence, this paper proposes a single-stage object detection model named Xnet to address this issue. The proposed method designs a deformable convolution backbone network used in the feature extraction stage. Compared to the standard convolution, it adds learnable parameters for dynamically analyzing the boundary and offset of the receptive field, rendering the model more adaptable to size variations within the same class. Moreover, this paper presents a novel anchor-free detector that classifies objects in feature images point-by-point, without relying on anchor boxes. Several experiments on the large remote sensing dataset DIOR challenging Xnet against other popular methods demonstrate that our method attains the best performance, surpassing by 4.7% on the mAP (mean average precision) metric.

Keywords: Anchor-free detector, Convolution neural network, Deformable convolution, Object detection, One stage detector, Remote sensing

1 Introduction

Object detection purposes include accurately classifying target categories and specifying position information in the input image. As a branch of computer vision, object detection plays a vital role in urban management, defense security, and environmental detection. Traditional object detection techniques typically extract feature vectors based on the input images' texture, brightness, and color and then, apply decision trees or support vector machines to regress the target information utilizing feature vectors. However, many complex objects, such as ports, airports, and highways, appearing in remote sensing images, do not exhibit consistent dimensions and shapes. Thus, the complex nature of the objects being detected prohibits handcrafted feature selection and

conventional detection methods from being accurate. Nevertheless, with the fast growth of deep learning in pattern recognition, more researchers are attempting to apply these techniques to various problems, such as image classification and instance segmentation. Indeed, deep learning techniques accommodate massive amounts of data without manual adjustment and fully extract expressive features in the image.

The effectiveness of deep learning methods in computer vision has led to much research about object detection in the remote sensing field. For instance, Li et al. [1] proposed a rotation-insensitive region proposal network and introduced a new anchor type based on Faster-RCNN that effectively deals with the rotation change of geospatial objects. To address the problem of appearance feature ambiguity and translation invariance, Zhong et al. [2] suggested a position-sensitive network that combines local and contextual properties. Liu et al. [3] replaced the traditional bounding box with a rotatable bounding box in the SSD framework, which is rotation invariant due to its ability to estimate object-orientation angles. Ding et al. [4] developed a full convolution neural network by dilated convolution and hard sample mining methods to achieve high-precision detection of small objects in remote sensing images. Dong et al. [5] proposed a new maximum suppression method, SIG-NMS, to reduce the error detection of small objects.

In order to detect objects from remote sensing images, Li et al. [6] designed an improved Transformer to aggregate the features of global spatial positions on multiple scales and used attention mechanism to adjust the data set differences with the pre-training model. Wang et al. [7] uses an improved Inception model to enhance the ability of small target extraction in shallow network and improves the pyramid structure in the model to enhance the effect of feature fusion. In addition, this method also compares the extracted features with candidate boxes with different horizontal-vertical ratios, so as to realize the detection of targets with different scales. Li et al. [8] proposed a novel Adjacent Context Coordination Network to explore the consistency of adjacent features in encoder and decoder structures. The network includes an ACCoM structure, which is used to activate salient regions in output features and coordinate multi-level features at the same time. Ye et al. [9] apply the stitcher to get an image containing objects with different scales, which can balance the ability of detecting multi-scale objects in the training process. In addition, this research is inspired by attention mechanism, and fuses spatial attention and channel attention in the model to obtain more representative feature information. Li et al. [10] proposed SeaNet, which includes MobileNet-V2 for feature extraction, DSMM for high-level dynamic semantic matching and ESAM for low-level feature edge self-alignment. This method firstly obtains the high-level features of the input remote sensing image and then, uses dynamic convolution to activate the position of salient objects in the high-level features. At the same time, in ESAM, cross-scale edge information extracted from low-level features is used for self-alignment and detail alignment. Yolov5-extract [11] algorithm was proposed by Tom et al. In this algorithm, the feature layer and prediction header with poor feature extraction ability in Yolov5 algorithm were removed, and a network mixed with Coordinate Attention and dilated convolution was integrated in the model, which optimized the extraction ability of object position and feature information at different scales. Dong et al. [12] proposed a remote sensing image target detection network based on FPN architecture. Considering

the variable shape and direction of remote sensing targets, this method uses deformable convolution to replace the horizontal connection in FPN to obtain the feature map with variable receptive field. In addition, this method also introduces several attention-based feature fusion modules to adaptively integrate the multi-level outputs of FPN, thus further realizing multi-scale target detection. Wang et al. [13] proposed a novel single-stage detector MSE-Net, which consists of a multi-scale enhanced network and a scale-invariant regression layer. First, MSE-Net provides multi-scale description enhancement by integrating Laplace kernel with fewer parallel multi-scale convolution. In addition, the model contains three different independent regression branches (corresponding to small, medium and large scales) in the regression layer, so that the default discrete scale bounding box covers the full-scale object information in the regression process.

Inspired by the current works, this paper develops a novel anchor-frame detection model named XNnet, to detect specific targets in remote sensing images. Compared to natural scenes, the targets' scale in remote sensing images varies significantly, posing essential impediments in the detection tasks. Thus, Xnet employs a deformable convolution strategy to construct a feature extraction network for a flexible receptive field during the training phase and adapt to the targets' shapes at each scale. Finally, the proposed method is challenged against other detection methods on the large remote sensing dataset DIOR and attains a superior performance. In conclusion, this study has the following contributions:

1. Proposing a single-stage, anchor-free object detection method. The feature extraction network generates high-quality output features by fusing multi-scale features, while the anchor-free detector classifies and regresses the output features point-by-point to predict the category and location information of the objects.
2. Due to the targets' variable shape in the dataset, the proposed feature extraction method implements deformable convolution to construct the feature extraction network. During the training phase, the deformable convolution dynamically modifies the range of the receptive fields to obtain more differentiated features.
3. The proposed method is compared to several popular detection methods on the DIOR dataset, revealing that our method achieves the best results among the single-stage methods. Specifically, our scheme significantly outperforms the competitor methods in mAP metrics and is comparable to the effectiveness of the multi-stage methods.

Overall, the contribution of this paper lies in the design of a feature extraction network incorporating deformable convolution and residual structure. The backbone network can dynamically adjust the range of receptive fields of the model according to the input feature, so that the model can capture abundant context information in multi-scale receptive fields. Another innovation of this work is that a novel anchor-free design structure is proposed, which not only eliminates the complex anchor boxes clustering in the previous anchor-based model in the preprocess stage, but also directly predicts the category and position of the target according to the feature pixels, which can better adapt to the detection of targets at different scales.

The structure of this article is as follows: in the first chapter, we introduce the innovation of this work, the existing problems of remote sensing image object detection and the methods proposed by this paper to solve the problems. In the second chapter, this paper introduces the research on object detection in satellite remote sensing images. The related research can be divided into two categories, one is single-stage, the other is multi-stage. In the third chapter, the proposed method is introduced in detail, including the improved feature extraction network, feature fusion module and the design of anchor-free detection head. In addition, from the perspective of clustering experiments, this chapter also analyzes the limitations of anchor-based methods. In the fourth and fifth chapters, this paper introduces in detail the design and the quantitative indicators related to the comparative experiment, and the analysis of the experimental results, as well as the experimental details in the training and verification stage.

2 Related work

Object detection based on convolutional neural networks involves single-stage and multi-stage methods.

2.1 Multi-stage object detection

Differentiating multi-stage methods necessitates regions from the region proposal network as candidates, as these methods' first stage generates a series of candidate region proposals that may contain objects. The second stage is to divide the candidate region proposals from the previous stage into objects and backgrounds and further adjust the coordinates of the detection boxes. The R-CNN method [14] is a representative multi-stage detection method, which applies a CNN (convolutional neural network) to extract the features of each proposal region and then, feeds the features into the support vector machine for further classification. Ren et al. [15] proposed the Faster-RCNN method, where the region proposal network (RPN) shares the convolutional parameters with the subsequent module. The RPN generates high-quality region proposals for the detection network through end-to-end training. An alternative solution is the feature pyramid network (FPN) [16], which introduces top-down paths into the feature extraction network. This strategy aims to improve the Faster-RCNN's feature extraction capability process by reducing the distance between the features. Besides, the quality of the whole feature can be improved by using the semantic information collected from the deep features.

Wu et al. [17] introduced the Double head RCNN and analyzed the effect of fully connected and convolution layers in object detection. Sun et al. [18] developed Sparse RCNN, a sparse prediction method that employs learnable region proposals to avoid a manual anchor box design and many-to-one label assignments. Cai et al. [19] suggested the Cascade RCNN method, which involves progressive training and resampling to ensure that all detectors have the appropriate proportions of the positive sample set, thereby reducing overfitting. Kim et al. [20] proposed the PAA (Probabilistic Anchor Assignment) detection method, a new anchor assignment strategy that adaptively divides the anchor boxes into positive and negative samples based on the model's learning state. However, the above methods rely on the region proposal network to extract all candidate region proposals, with multi-stage methods excluding irrelevant regions before classification, thereby improving detection accuracy.

2.2 Single-stage object detection

The single-stage target detection strategy does not need to generate candidate region proposals compared to the multi-stage approach. A representative approach is the YOLO [21, 22] series, which uses the output features of the CNN backbone and pre-set anchor boxes to predict the detection boxes directly and provide the corresponding class probabilities in the whole image. Compared with YOLO, SSD [23] achieves better results in locating small-sized objects by introducing multi-scale features and default boxes with multiple groups. Moreover, Lin et al. [24] proposed RetinaNet, which utilizes a hard sample mining approach to alleviate unbalanced foreground and background samples in the object detection task. Li et al. [25] suggested the Trident Network detection method, which relies on a multi-branch structure with shared parameters but different receptive fields. Each branch trains and predicts objects within a specific range of size. Tan et al. [26] introduced EfficientDet, a method for dynamically scaling the model's width, dimensionality, and depth. In recent years, research on anchor-free detection methods has gained significant attention. For instance, Law et al. [27] developed an anchor-free detection method, Corner Net, in which a pair of key points (upper left and lower right corners) are used as detection targets and then, matched in pairs in the same detection box by a new grouping method. Duan et al. [28] suggested the CenterNet detection model that uses key point estimation to find the target's centroid and regress other object attributes, i.e., location and orientation. Ge et al. [29] proposed the anchor-free approach YOLOX, which decouples the network's detection and classification heads. This method uses SimOTA to match the predicted detection boxes with the ground truth in the training phase.

However, high-precision object detection has always been an essential task for optical remote sensing image processing, and there is still ample room for development. In the proposed model, the feature extraction network relies on deformable convolution [30] to adaptively capture the target-related feature information. Simultaneously, an anchor-free detector classifies and regresses target information in the multi-level output feature. Based on these enhancements, this method increases the object detection precision in remote sensing images and reduces misidentification and leakage identification.

3 Methods

The proposed model comprises the feature extraction network, the feature fusion module, and the detection module (Fig. 1). Residual networks [31] and deformable convolutions are utilized to construct the feature extraction module, which extracts feature maps at varying levels. The feature fusion module uses a feature pyramid network to fuse essential information of multi-level features, while deformable convolution improves the output features and adjusts the receptive field area dynamically during training, accommodating geometric changes in target objects. The multi-stage output features from the feature extraction network are input into the module for anchor-free detection. The detection module predicts object categories and regresses locations where the gradient value in the feature image is greater than zero. The subsequent sections analyze each module within our method.

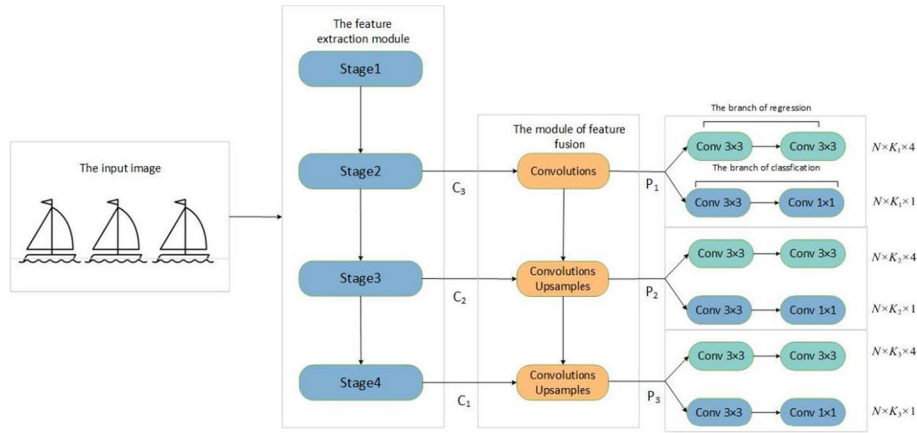


Fig. 1 The overall structure of the proposed model, where N is the batch size and K is the number of detection boxes

3.1 The feature extraction network

3.1.1 Deformable convolution

Equation (1) presents the two-dimensional convolution formula, where K represents the convolution operation, and the size of the input features I is $m \times n$. Obtaining the complete feature information of the target is a crucial step for subsequent model classification, with Dai et al. [30] introducing the limitation of the normal convolution and explaining why the network meets the challenges when detecting complex objects.

$$O(i, j) = \sum_m \sum_n I(i + m, j + n) \cdot K(m, n), \quad (1)$$

A normal convolution extracts features within the radius of the receptive field. However, the limited size and shape of the receptive field might lead to incomplete target object features, affecting the accuracy in the final detection stage. Especially most objects in remote sensing images have ambiguous bounds and various scales.

As shown in Eq. (2), the deformable convolution adds two learnable variables, x and y , to the normal convolution, assisting the network to flexibly sample features during the training stage.

$$O(i, j) = \sum_m \sum_n I(x + m + \Delta x, j + n + \Delta y) \cdot K(m, n), \quad (2)$$

In addition, Fig. 2 visually compares the receptive field between the normal and deformable convolution, highlighting that deformable convolution has a broader sampling range and better meets the target distribution. Hence, the deformable convolution allows the network to obtain complete and precise feature information.

3.1.2 The structure of the feature extraction network

The feature extraction network in the proposed model substitutes some of the normal convolutions with deformable convolutions based on ResNet [31], which affords an appealing performance and is frequently employed in various visual tasks. A

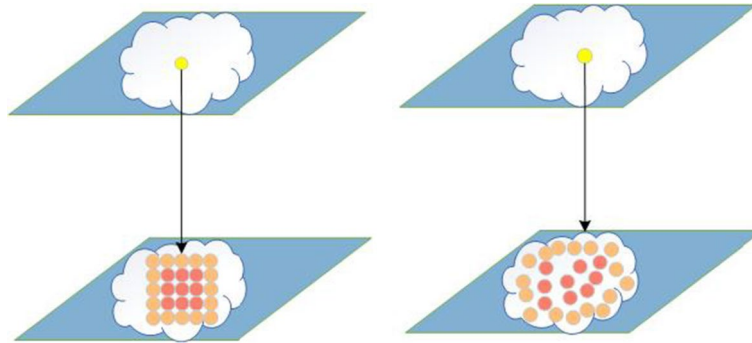


Fig. 2 Visual comparison of the deformable convolution (right) and the normal convolution (left)

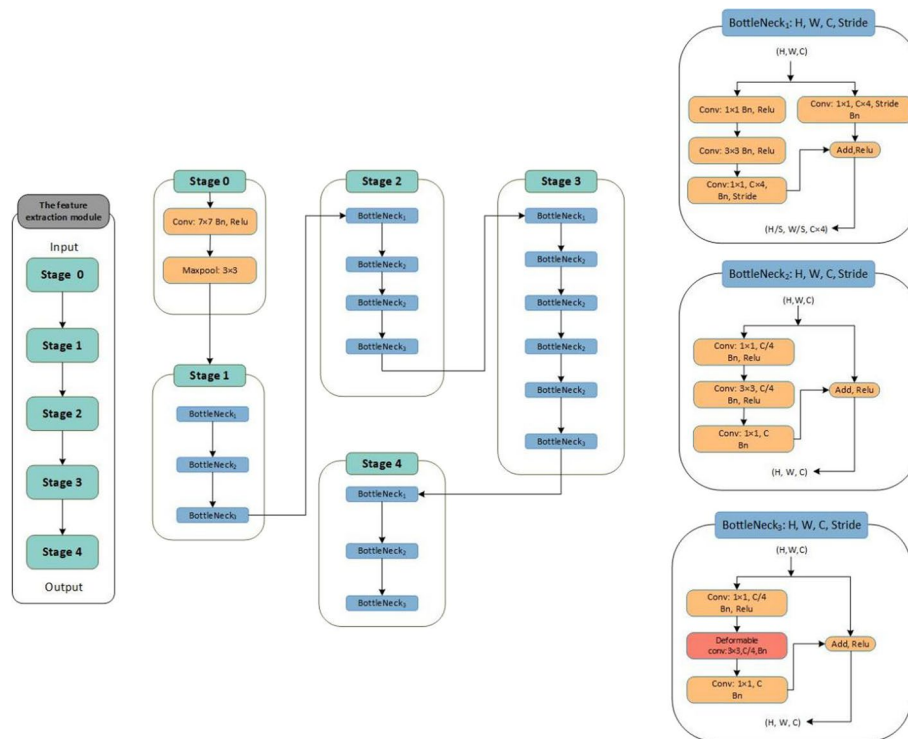


Fig. 3 The structure of the feature extraction module

deformable convolution involves two additional learnable parameters enabling the network to adjust the receptive field and improve the network's ability to adapt to target scale changes. As depicted in Fig. 3, the feature extraction network comprises five stages, each composed of three distinct bottle structures. Each bottle structure encompasses multiple convolution layers with a 3×3 kernel size and residual connections.

The difference between the three bottleneck structures is whether a 1×1 convolution adjusts the feature channel and whether the deformable convolution is supplanted for the normal convolution at the end of the bottleneck structure. Figure 3 highlights that the deformable convolution exists only in the main branch of

BottleNeck3 due to computational efficiency. The input size of the feature extraction module is $w \times h \times c$, where w , h , and c are the width, height, and channel, respectively. The output feature size is $\frac{w}{2^i} \times \frac{h}{2^i} \times (c \times 2^i)$, where i is the stage of the output feature. As illustrated in Fig. 1, the outputs of different stages are input into the next module.

3.2 The feature fusion module

Shallow features in convolutional neural networks commonly comprise more positional information than deep features, whereas deep features contain more semantic information due to their larger receptive fields. Moreover, multiple downsampling imposes the target's feature information loss, particularly weakening the small objects' category and location information. Consequently, the proposed model employs the feature fusion module to enhance the detection effect of the small-scale target.

According to Fig. 4, the feature fusion module utilizes pyramid structures to connect features directly at various levels. Moreover, the deep and shallow features are fused in the fusion module. The upsampling operation forces the deep features to have the same size as their adjacent shallow features, and then, the module concatenates the adjacent features and feeds them to the subsequent stage. If the channel numbers of two concatenated feature maps are unequal, the module first applies a 1×1 convolution operation to adjust both channel numbers. The feature map C_1 , C_2 , and C_3 , generated by the feature extraction module, are input to the feature fusion module, and the i -th output feature P_i has the same resolution as the input feature C_i . Therefore, the anchor-free detector in the subsequent stage obtains richer image features by enhancing its ability to detect small objects.

3.3 Anchor-free detector

We further describe the uniqueness of the anchor-free detector head in the following contents. In general, object detection networks with anchor frames have a great impact on detecting some targets of different sizes, and more preset anchor boxes are needed to get higher recall during training, so this leads to an additional need to calculate the maximum intersection ratio between anchor boxes and ground truth during training.

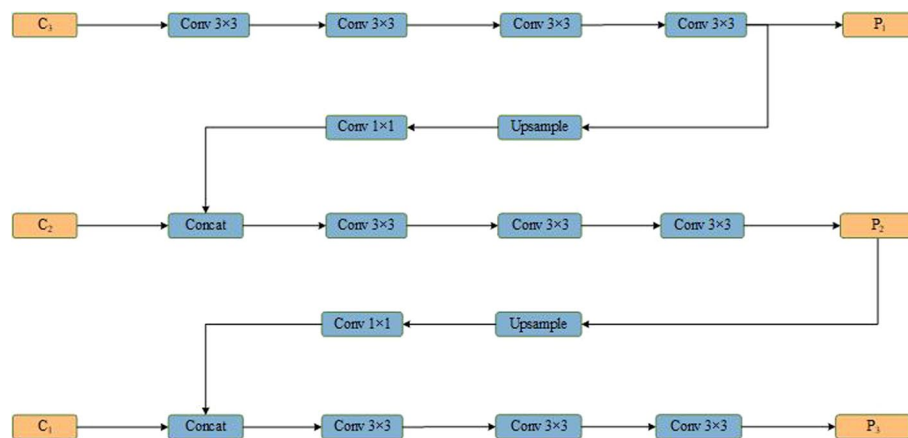


Fig. 4 The structure of the feature fusion module

In addition, since most of the anchor boxes are negative samples, there is also the problem of sample imbalance. However, in the anchor-free detection model proposed in this paper, the calculation of IOU between anchor box and ground truth is avoided, and the good detection effect is maintained. According to the description above, the Xnet method proposed in this paper introduces multi-scale pixel-by-pixel regression for object detection while using an anchor-free box detection head. The scale of targets regressed on different scale features is varied, and for a certain layer, targets that do not satisfy the regression size of that layer are not processed further, which also alleviates the duality due to target overlap to some extent.

Three features from the feature extraction network are input into the feature fusion module to generate the output features $P_i, i \in (1, 2, 3)$ in three varying scales. In contrast to anchor-based methods, the anchor-free detector directly predicts the object's category and location information in multi-stage output features. To investigate the relationship between feature gradients and target distribution, we employ the Gram-Cam [32] method to generate a gradient map of the input image. In Fig. 5, the highlighted regions represent pixels with strong gradients that overlap heavily with the target objects. Spurred by this finding, we eliminate the interference of irrelevant regions by employing Xnet, which only performs prediction in regions where the gradients are greater than zero.

As shown in Fig. 6, the detector regresses a certain point (x, y) to obtain a four-dimensional vector $t = (\hat{l}, \hat{h}, \hat{r}, \hat{b})$ and its corresponding label \hat{c} , where $(\hat{l}, \hat{h}, \hat{r}, \hat{b})$ is the distance between point (x, y) and the boundary of the detection box and \hat{c} denotes the classification of the detection box. For point (x, y) in the feature image, $(x - \hat{l}, y - \hat{r})$ on the upper-left corner and $(x + \hat{r}, y + \hat{b})$ on the lower-right corner constitute the coordinates of the detection box, respectively.

Each point satisfying this requirement has detection boxes of various scales for the input feature images. In other words, the same point has a detection box for the feature image at different levels. The coordinate in P_i must be multiplied by 2^{i+1} to match the size of the original images. The anchor-free detection module comprises two fully convolutional branches, where the classification branch predicts the probability that a point belongs to the category, and the regression branch predicts the center location and offset of the bounding box.

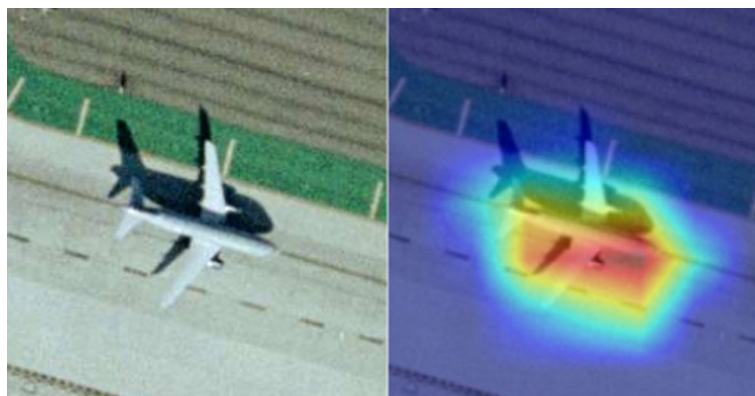


Fig. 5 The gradient map of the input image

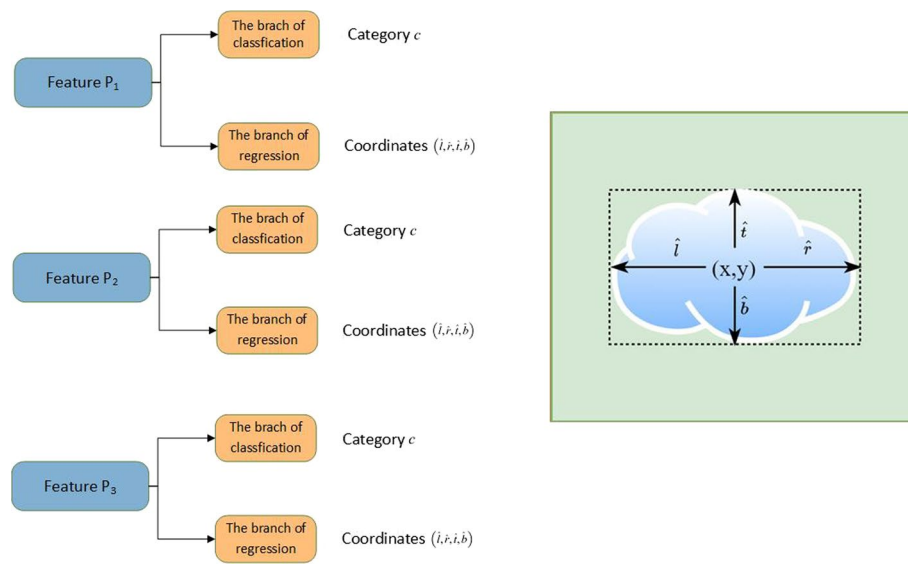


Fig. 6 The structure of the anchor-free detector

Table 1 The IoU between the preset anchor boxes and the ground truths

Number of groups for clustering	Mean IoU
3	0.4434
4	0.4703
5	0.5060
6	0.5190
7	0.5583
8	0.5706

3.4 The analysis of the anchor-based detector

This paper analyzes the performance of the anchor-based methods by clustering anchor boxes according to their width and height using the KMeans++ algorithm [33]. Table 1 reports that the IoU between limited anchor boxes and the ground truth is small, with a decreasing trend as the number of cluster groups rises. This indicates that in the initial stage, preset anchor boxes cannot adequately accommodate the sizes of the detected objects, necessitating substantial position adjustments during the training phase of the anchor-based method. Our novel anchor-free detector, verified in a comparative experiment, enables our method to resolve this issue.

4 Loss function

In Eq. (3), the cross-entropy loss assesses the reliability of the classification results. If the model classifies the results more precisely, the loss value tends to decline. In Eq. 3, y_{ic} denotes the i -th sample that belongs to category c , and p_{ic} is the predicted probability it belongs to category c .

$$L_1 = -\frac{1}{N} \sum_i \sum_{c=1}^M y_{ic} \log(p_{ic}), \quad (3)$$

The IoU loss in Eq. (4) represents the degree to which the detection and ground truth boxes overlap and is primarily employed to evaluate the precision of the proposed model's regression section. In this loss function, A represents the ground truth coordinate, B is the detection box coordinate, and $P_{xy} > 0$ indicates that the model only regresses the detection boxes for regions where the feature gradient is greater than zero, ignoring all other regions.

$$L_2 = N_{\text{pos}} \sum_{\text{reg}} p_{xy>0} L_{\text{reg}}(b_{xy}, \hat{b}_{xy}), \quad (4)$$

The loss function in Eq. (5) consists of the two loss functions listed above, where N_{pos} is the number of images sampled during the training phase and λ is applied to balance the weight of the classification. In the subsequent comparison experiments, the IoU loss is set to 1.

$$L = L_1 + \lambda \cdot L_2, \quad (5)$$

5 Experiment

5.1 The DIOR dataset

We demonstrate the proposed model's efficacy by challenging it against current methods on the large DIOR remote sensing image dataset. DIOR [34] comprises 23,463 remote sensing images covering 20 categories of ordinary objects. The images have a fixed resolution of 0.5–30 m and a fixed size of 800×800 . Figure 7 depicts the categories distribution, which is diverse in data quantity and object categories. The target area in the image varies from 100 to 4×10^4 , including 20 categories such as aircraft, airports, baseball fields, bridges, chimneys, and dams. To ensure that the object class during training

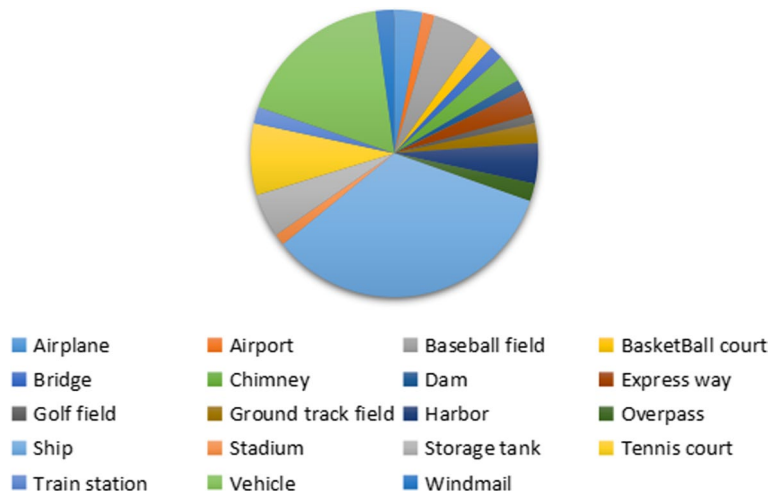


Fig. 7 Class distribution in the DIOR dataset

remains balanced, the experiments exploit the officially divided training and validation sets.

5.2 Training and inference details

Our method is challenged against current deep learning methods under the default configuration of the Paddle framework, i.e., all network parameters are initialized randomly and retrained on the experimental dataset. During the training stage, all methods adjusted the input resolution to 800×800 . Moreover, the comparison experiment countered overfitting through image augmentation techniques such as horizontal flipping, scaling, clipping, and color vibration. The batch size for each iteration is eight, and the initial learning rate is 1×10^{-3} . During the training stage, 36,000 iterations are performed. When the model's loss fails to decrease after 5000 iterations, the learning rate reduces to 1/10 of its previous value until the minimum learning rate 1×10^{-5} is reached. Only the weights with the minimum loss are saved during the validation phase. The model obtains the detection boxes of the input image during the prediction phase, with the maximum suppression method neglecting the redundant detection boxes to provide the final output.

All trials are implemented on an Intel Xeon Gold 6330 CPU with 32 GB of memory and an NVIDIA 3090 graphics card with 12 GB of video memory.

5.3 Performance evaluation metrics

Object detection models are typically evaluated based on the accuracy, precision, recall, and specificity performance metrics, with different metrics utilized based on the application and scenario. This paper employs the mAP (mean average precision) metric involving precision and recall, defined as:

$$\text{Precision} = \frac{TP}{TP + FP}, \quad (6)$$

$$\text{Recall} = \frac{TP}{TP + FN}, \quad (7)$$

where TP, FN, FP, and TN are defined as presented in Table 2.

Equation (8) defines the method to plot the PR curve, where the X-axis denotes the prediction accuracy and Y-axis is the recall. Calculating the area within the curve and the coordinate axis determines the predicted AP value of one category in the dataset. The mAP value is the average of the AP values of all categories.

Table 2 The definitions of TP, FN, FP, and TN

The ground truth	Prediction value	
	Positive samples	Negative samples
Positive samples	TP	FN
Negative samples	FP	TN

$$AP = \int_0^1 P(R) dR, \quad (8)$$

The Intersection over Union (IoU) is considered from 0.5 to 0.95 when evaluating the model's prediction performance. We calculate the AP values corresponding to different IoUs and then, average them to obtain $mAP@0.5:0.95$. In addition, the mAP_{small} , mAP_{mid} , and mAP_{large} metrics represent the AP values of varying-sized detection targets, where small targets are smaller than 1024 pixels, medium targets are between 1024 and 4096 pixels, and large targets are greater than 4096 pixels.

5.4 Analysis of the experimental results

As reported in Table 3, the mAP value of the YOLOV4 method is only 24.5%. As a representative anchor-based object detection method, the performance of YOLOV4 in the DIOR data set lagged significantly with the anchor-free methods. Therefore, the anchor-free method is more appropriate for datasets involving a wide distribution of object size and categories. In the $mAP@0.5:0.95$ and mAP_{small} metrics, the proposed method deviated slightly from the Cascade-RCNN model, which is a multi-stage method, presenting only a 0.3% gap in the capacity to detect small objects.

Therefore, the single-stage Xnet method can nearly achieve a similar detection performance compared with the Cascade-RCNN method. In contrast to other prevalent methods, our model achieves the optimal performance on the test dataset, while $mAP@0.5:0.95$ increases by 4.7% compared to the suboptimal YOLOX method. According to the feature point distribution, CenterNet [27] can regress the size and location of the detection box. However, the feature points of some targets in the feature maps will be compressed to the same location during downsampling, resulting in the overlapping problem.

Xnet obtains more discriminative features and location information of multi-level features at various scales to prevent the occurrence of the above problem. Consequently, compared to CenterNet, our technique increases $mAP@0.5:0.95$ and mAP_{small} by 7.7% and 2.1%, respectively. The suggested method obtains the complete output features with less information redundancy through deformable convolution than the YOLOX method, thereby reducing the interference of useless information in object detection. Accordingly, our method affords a 4.7% higher mAP in the test dataset, while the mAP_{small} is slightly lower than the YOLOX method by 0.7%.

According to the above analysis, additional annotation information is not required in the proposed model. The deformable convolution design in the feature extraction












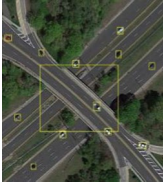







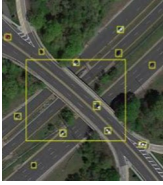
Table 3 Metrics in the comparative experiment

Methods	$mAP@.5:.95$	mAP_{small}	mAP_{mid}	mAP_{large}
Ours	0.498	0.147	0.397	0.689
CenterNet	0.421	0.124	0.363	0.586
YOLOV4	0.245	0.051	0.206	0.351
YOLOX	0.451	0.154	0.386	0.602
Cascade RCNN (Multi-stage)	0.517	0.150	0.425	0.705

network and the anchor-free detection are applied to fuse the output feature maps of different scales effectively and mitigate the sizable differences within the same category to enhance the performance of object detection in remote sensing images.

This section demonstrated this method’s efficacy by comparing and analyzing the experimental evaluation metrics. Similarly, as demonstrated in Table 4, the detection performance of Xnet is superior to the competitor methods. For instance, Example C in Table 4 reveals that only our method detected the ships of all sizes and the ports where they are docked. In Example B, no errors or omissions are made in detecting the aircraft with dense stops. Hence, Xnet can adapt to detecting both types with significant size differences within the same image (Example D). As demonstrated by the visual examples, Xnet achieves the same excellent effect as the multi-stage detection method, demonstrating its detection efficacy.

Table 4 Comparison of the visualization results

Methods	Example			
	A	B	C	D
Ours				
CenterNet [27]				
YOLOV4 [35]				
Cascade RCNN [19]				
YOLOX [29]				

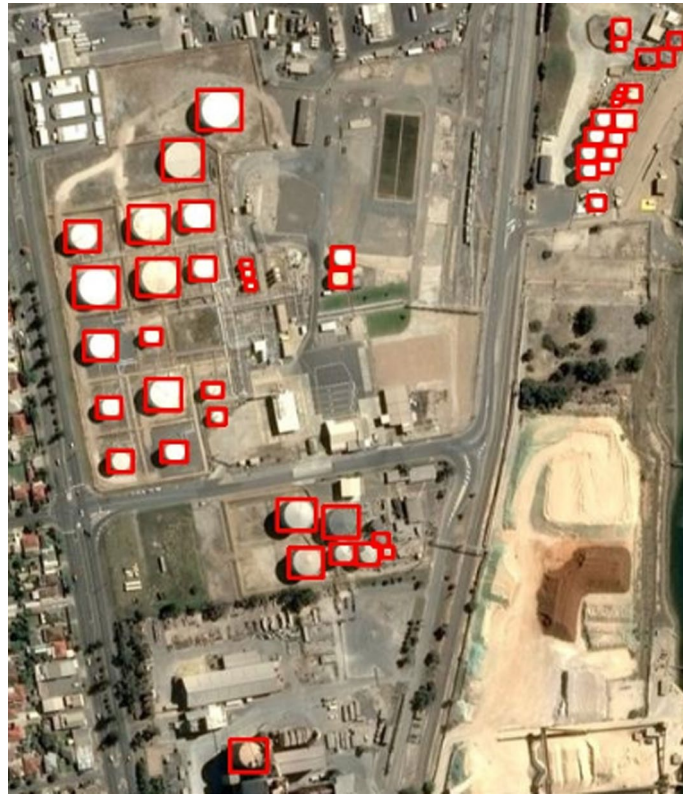


Fig. 8 Sample in the ADSI dataset

Table 5 Metrics in the comparative experiment

Methods	mAP@.5:.95	mAPsmall	mAPmid	mAPlarge
Ours	0.513	0.137	0.407	0.699
CenterNet	0.445	0.122	0.353	0.593
YOLOV4	0.411	0.119	0.322	0.412
YOLOX	0.427	0.134	0.397	0.631
Cascade RCNN (Multi-stage)	0.501	0.150	0.411	0.715

ADSI (Airbus Defense and Space Intelligence) is an oil storage tank detection dataset provided by Airbus. Storage areas for large oil tanks are common in manufacturing and government facilities, and are critical infrastructure in the industry.

Deep learning uses aerial or remote sensing images to detect oil tanks, which can get the number, type and status information of oil tanks in the current storage area, thus preventing oil tanks from overturning, leaking and other events. The following figure is a sample in the ADSI dataset, and the distribution of oil tanks in this scenario as Fig. 8 is in the red rectangular box.

As shown in Table 5, the mAP value of the YOLOX method is 43.1%, which is also a method for detecting targets without anchor boxes, the detection effect of YOLOX method is far inferior to the another anchor-free method Xnet proposed in this paper.

Overall, the results of this comparative experiment are similar to the previous comparative experiments conducted on the DIOR dataset, and the anchor-free detection methods also exhibit better detection results than the anchor-based detection methods. Among the anchor-free detection methods, the method with the lowest mAP 42.7% is better than any one of the methods with anchor boxes. The following Fig. 9 shows the actual detection effect of the method proposed in this paper, although there were numerous oil tanks to be detected, the Xnet method was able to detect all of them in the scene.

5.5 The ablation experiments

The ablation experiments aim to confirm the influence of deformable convolution on object detection. To ensure experimental integrity, the super parameter setting and training strategy of the ablation experiment are identical to the comparative experiment. The effects of the ablation experiment are listed in Table 5, which highlights that when the structure of the anchor-free detector in the proposed method is unaltered, the design of the deformable convolution in the feature extraction network improves the overall model's detection effect. There is a 2.2% improvement in the mAP@0.5:0.95 metric and a 1.2% increase in the mAP_{small} metric. To test the effect of deformable convolution on various feature extraction networks, we create a new control group utilizing another backbone network, while all other experimental settings remain unchanged. As indicated by the experiment metrics, the deformable convolution significantly impacts



Fig. 9 The actual detection effect of the method proposed in this paper

Table 6 Metrics in the ablation experiment

Methods	mAP@.5:.95	mAP _{small}	mAP _{mid}	mAP _{large}
Resnet50 with deformable convolution	0.498	0.147	0.397	0.689
Resnet18 with deformable convolution	0.373	0.102	0.285	0.560
Resnet50 without deformable convolution	0.476	0.135	0.365	0.671
Resnet18 without deformable convolution	0.370	0.096	0.281	0.553

ResNet18, attaining a 0.3% improvement in mAP@0.5:0.95 and a 0.6% improvement in mAP_{small}. According to the findings of the experiments, deformable convolution can improve the accuracy of an object detection task, accomplishing a superior outcome.

The method described in this paper achieves the same improvement effect with different feature extraction networks, exemplifying that this method can augment the object detection effect of other feature extraction networks, as shown Table 6.

6 Conclusion

The challenge in object detection of remote sensing images arose from the considerable proportion of small-size objects and the differences within the same type of objects, leading to the ineffective detection effect of some current multi-stage and single-stage methods. Regarding the anchor-based method, the anchor boxes set manually or based on clustering cannot cover the size distribution of all data set objects.

To address the issues above, we incorporate a deformable convolution for feature extraction instead of exploiting standard convolution, providing the network with a more adaptable receptive field. In addition, our anchor-free detector directly predicts the detected objects on the multi-level feature map without requiring preset anchor boxes to adjust the position, thereby overcoming the accuracy degradation caused by the dense distribution of anchor boxes and the sample imbalance in the anchor-based method. The comparative experiments demonstrate that although our scheme adopts a single-stage detection strategy, it affords an IoU of 49.8%, comparable to multi-stage models, thus verifying our method's efficacy. Future research will investigate reducing the model's parameters and applying our model in a scenario with limited computational resources.

This research additionally conducted a comparison experiment on the ADSI dataset. The results of this comparison experiment show that for small object detection in remote sensing images, the anchor-free detection model is more advantageous in terms of detection accuracy. In addition, XNet achieves the best detection results among the anchor-free methods in the comparison experiment with the feature extraction capability of the backbone network incorporating deformable convolution and the multi-layer feature fusion network, as well as the unique anchor-free detection head design.

Acknowledgements

This work is supported by Natural Science Foundation of Fujian Province under Grant Nos. 2022J01130395, 2019J05123).

Authors' information

Shenshen Fu, received his M.S. degree in Computer Science from Zhejiang Wanli University in 2019. He is working toward a master's degree with the school of computer and information engineering, Xiamen University of Technology, Xiamen, China. His research interests include image processing and computer vision. Yifan He, received his M.S. degree (cum laude) and Ph.D. degree in electrical engineering from the Eindhoven University of Technology (TU/e), The Netherlands, in 2008 and 2013, respectively. He is now the CTO of Reconova Co. Ltd. His

current research interests include deep neural networks, low-power computing systems, and computer architectures of AI chips. He has published over 50 articles in international journals and conferences.

Xiaofeng Du, received her M.S. in Computer Science from Shanghai University, China, in 2006 and his Ph.D. in Computer Engineering from Xiamen University, China, in 2010. From 2010 to 2012, she was a postdoctoral researcher at Xiamen University. Since 2012, he has joined the School of Computer and Information Engineering, Xiamen University of Technology. Her research interests include image processing and computer vision.

Yi Zhu, received a B.S. degree in Computer Science from the Nanjing Institute of Technology. He is employed by Guodian Nanjing Automation Co. Ltd. and ABB (China) Ltd. Yi Zhu has designed several standardized protocols for electrical communications. He is working toward a master's degree at the Xiamen University of Technology. His research focuses on developing large-scale software for image processing and electrical automation.

Author contributions

All authors read and approved the final manuscript.

Availability of data and materials

The data that support the findings of this study are available from author, Yi Zhu, upon reasonable request.

Declarations

Competing interests

The authors declare no conflict of interest.

Received: 14 November 2022 Accepted: 17 April 2023

Published online: 09 May 2023

References

1. K. Li, G. Cheng, S. Bu, X. You, Rotation-insensitive and context-augmented object detection in remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **56**, 2337–2348 (2018). <https://doi.org/10.1109/TGRS.2017.2778300>
2. Y. Zhong, X. Han, L. Zhang, Multi-class geospatial object detection based on a position-sensitive balancing framework for high spatial resolution remote sensing imagery. *ISPRS J. Photogramm. Remote Sens.* **138**, 281–294 (2018). <https://doi.org/10.1016/j.isprsjprs.2018.02.014>
3. W. Liu, L. Ma, H. Chen, Arbitrary-oriented ship detection framework in optical remote-sensing images. *IEEE Geosci. Remote Sens. Lett.* **15**, 937–941 (2018). <https://doi.org/10.1109/LGRS.2018.2813094>
4. P. Ding, Y. Zhang, W.J. Deng, P. Jia, A. Kuijper, A light and faster regional convolutional neural network for object detection in optical remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **141**, 208–218 (2018). <https://doi.org/10.1016/j.isprsjprs.2018.05.005>
5. R. Dong, D. Xu, J. Zhao, L. Jiao, J. An, Sig-NMS-based faster R-CNN combining transfer learning for small target detection in VHR optical remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **57**, 8534–8545 (2019). <https://doi.org/10.1109/TGRS.2019.2921396>
6. Q. Li, Y. Chen, Y. Zeng, Transformer with transfer CNN for remote-sensing-image object detection. *Remote Sens.* **14**(4), 984 (2022)
7. L. Wang et al., *A Novel Deep Learning-Based Single Shot Multibox Detector Model for Object Detection in Optical Remote Sensing Images* (Wiley Online Library, New York, 2022)
8. G. Li, Z. Liu, D. Zeng, W. Lin, H. Ling, Adjacent context coordination network for salient object detection in optical remote sensing images. *IEEE Trans. Cybern.* **53**(1), 526–538 (2022)
9. Y. Ye et al., An adaptive attention fusion mechanism convolutional network for object detection in remote sensing images. *Remote Sens.* **14**(3), 516 (2022)
10. G. Li, Z. Liu, X. Zhang, W. Lin, Lightweight salient object detection in optical remote sensing images via semantic matching and edge alignment. *IEEE Trans. Geosci. Remote Sens.* **61**, 1–11 (2023)
11. Z. Liu, Y. Gao, Q. Du, M. Chen, W. Lv, YOLO-extract: improved YOLOv5 for aircraft object detection in remote sensing images. *IEEE Access* **11**, 1742–1751 (2023)
12. X. Dong, Y. Qin, Y. Gao, R. Fu, S. Liu, Y. Ye, Attention-based multi-level feature fusion for object detection in remote sensing images. *Remote Sens.* **14**(15), 3735 (2022)
13. G. Wang, Y. Zhuang, H. Chen, X. Liu, Q. Sang, FSoD-Net: full-scale object detection from optical remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **PP**(99), 1–18 (2021)
14. R. Girshick, J. Donahue, J. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014* (2014), pp. 580–587
15. S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**, 1137–1149 (2017). <https://doi.org/10.1109/TPAMI.2016.2577031>
16. T.Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021* (2017), pp. 936–944
17. Y. Wu, Y. Chen, L. Yuan, Z. Liu, L. Wang, H. Li, Y. Fu, Rethinking classification and localization for object detection, in *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020* (2020), pp. 10183–10192

18. P. Sun, R. Zhang, Y. Jiang, T. Kong, C. Xu, W. Zhan, M. Tomizuka, L. Li, Z. Yuan, C. Wang et al., Sparse R-CNN: end-to-end object detection with learnable proposals, in *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, 20–25 June 2021 (2021), pp. 14449–14458
19. Z. Cai, N. Vasconcelos, Cascade R-CNN: delving into high quality object detection, in *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, 18–23 June 2018 (2018), pp. 6154–6162
20. K. Kim, H.S. Lee, Probabilistic anchor assignment with IoUprediction for object detection, in *Proceedings of the Computer Vision—ECCV 2020*, Glasgow, UK, 23–28 August 2020 (2020), pp. 355–371
21. J. Redmon, A. Farhadi, YOLOv3: An Incremental Improvement (2018). [arXiv:1804.02767](https://arxiv.org/abs/1804.02767)
22. J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: unified, real-time object detection, in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 27–30 June 2016 (2016), pp. 779–788
23. W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.Y. Fu, A.C. Berg, SSD: single shot multibox detector, in *Proceedings of the Computer Vision—ECCV 2016*, Amsterdam, The Netherlands, 11–14 October 2016 (2016), pp. 21–37
24. T.Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **42**, 318–327 (2020). <https://doi.org/10.1109/TPAMI.2018.2858826>
25. Y. Li, Y. Chen, N. Wang, Z.X. Zhang, Scale-aware trident networks for object detection, in *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South), 27 October–2 November 2019 (2019), pp. 6053–6062
26. M. Tan, R. Pang, Q.V. Le, EfficientDet: scalable and efficient object detection, in *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 13–19 June 2020 (2020), pp. 10778–10787
27. X. Zhou, J. Zhuo, P. Krähenbühl, Bottom-up object detection by grouping extreme and center points, in *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 15–20 June 2019 (2019), pp. 850–859
28. K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, Q. Tian, CenterNet: keypoint triplets for object detection, in *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South), 27 October–2 November 2019 (2019), pp. 6568–6577
29. Z. Ge, S. Liu, F. Wang, Z. Li, J. Sun, YOLOX: Exceeding YOLO Series in 2021 (2021). [arXiv:2107.08430](https://arxiv.org/abs/2107.08430)
30. J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, Y. Wei, Deformable convolutional networks, in *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 22–29 October 2017 (2017), pp. 764–773
31. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *Proceedings of the Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 27–30 June 2016 (2016), pp. 770–778
32. R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: visual explanations from deep networks via gradient-based localization, in *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 22–29 October 2017 (2017), pp. 618–626
33. D. Arthur, S. Vassilvitskii, *k-Means++: The Advantages of Careful Seeding*, 778 (Stanford InfoLab, Stanford, 2008)
34. K. Li, G. Wan, G. Cheng, L. Meng, J. Han, Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS J. Photogramm. Remote Sens.* **159**, 296–307 (2020). <https://doi.org/10.1016/j.isprsjprs.2019.11.023>
35. A. Bochkovskiy, C.-Y. Wang, H.-Y.M. Liao, YOLOv4: Optimal Speed and Accuracy of Object Detection (2020). [arXiv:2004.10934](https://arxiv.org/abs/2004.10934)

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► [springeropen.com](https://www.springeropen.com)