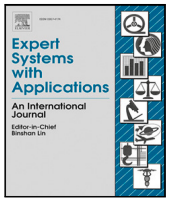
Expert Systems With Applications 225 (2023) 120068

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



FDLR-Net: A feature decoupling and localization refinement network for object detection in remote sensing images

Jinsheng Xiao [a], Yuntao Yao [a], Jian Zhou [b,*], Haowen Guo [a], Qiuze Yu [a], Yuan-Fang Wang [c]

[a] School of Electronic Information, Wuhan University, Wuhan, 430072, China
[b] State Key Laboratory of Information Engineering in Surveying, mapping and Remote Sensing, Wuhan University, Wuhan, 430072, China
[c] Department of Computer Science University of California, Santa Barbara, 93106, CA, United States of America

ARTICLE INFO

Keywords:
Object detection
Remote sensing
Arbitrary orientation

ABSTRACT

Object detection in remote sensing images is a critical task in computer vision. Often times in remote sensing images, objects are highly variable in scale and have arbitrary orientation, which renders spatial alignment between anchor boxes and objects challenging in the object detection task. In this paper, a feature decoupling and localization refinement network is suggested as a solution to this issue. Specifically, a bidirectional feature fusion module (BFFM) is devised to construct a multi-scale feature pyramid for detecting objects at different scales. A feature decoupling module (FDM) is devised which utilizes the fusion of spatial attention and channel attention, as well as different attention functions to generate features specifically tuned for regression and classification, that are used to guide more accurate localization and classification. Further, a localization refinement module (LRM) is designed to automatically optimize the anchor box parameters to achieve spatial alignment of the anchor box and the object regression feature. In this way, the FDM and LRM are cascaded to achieve more accurate localization. Experimental results on two open access datasets, DOTA and HRSC2016, show that the performance of FDLR-Net is state-of-the-art, with mAP reaching 73.08% and 89.4%, respectively.

1. Introduction

Optical remote sensing images are widely used for land and ocean observation due to their rich information and high spatial resolution, and play a vital role in land resource management (Jalayer et al., 2022), urban climate research (Wahla et al., 2022), and other fields. Object detection is one of the essential technologies of remote sensing image processing. Accurate location and classification for specified objects in remote sensing images are important tasks in areas such as sea surveillance and urban planning. Different from images taken, say, by a mobile phone, the view of remote sensing images is top-down, thus the objects in them have arbitrary orientation resulting in many horizontal object detection algorithms (Bochkovskiy et al., 2020; Carion et al., 2020; Xiao et al., 2023) that perform well in natural scenes often fail to obtain ideal detection results in remote sensing images. In addition, because of the high resolution of remote sensing images, the objects therein may appear to be small in size or vary in a wide range of scales, which further increases the difficulty of object detection.

Inspired by works on text detection (Jiang et al., 2017; Ma et al., 2018), researchers start to use rotation bounding boxes for remote sensing image object detection. Most of the existing rotation object detection methods solve the problem of arbitrary orientation and multi-scales of the object by presetting a certain number of anchor boxes with different scales, aspect ratios and angles, such as methods (Azimi et al., 2018), R^2PN (Zhang et al., 2018) and BoxNet (Neves et al., 2020). However, a large number of anchor box parameters causes an increase in computation. RoI-Transformer (Ding et al., 2019) avoids this problem by learning the transformation from horizontal Region of Interests (RoIs) to rotational RoIs. To better adapt to the multi-scale variation of targets, methods such as CAD-Net (Zhang et al., 2019), (Xiao et al., 2022) and Info-FPN (Chen et al., 2023) utilize contextual information or multiscale representation to construct more powerful feature representations. To improve the detection of small objects, SCRDet (Yang et al., 2019) refines sampling by adjusting the stride of the anchor to reduce the effect of insufficient training samples and imbalance. In addition, the paper also proposes an improved smooth L1 loss that can solve the boundary problem of the rotation bounding box. To align the rotation anchor box with the object region more accurately, R^3Det (Yang et al., 2021) designs a feature refinement module to perform regression from coarse to fine granularity by obtaining more accurate features.

* Corresponding author.
E-mail addresses: xiaojs@whu.edu.cn (J. Xiao), dolphin_tao@whu.edu.cn (Y. Yao), jianzhou@whu.edu.cn (J. Zhou), guohw2020@whu.edu.cn (H. Guo), yuhenry007@whu.edu.cn (Q. Yu), yfwang@cs.ucsb.edu (Y.-F. Wang).

https://doi.org/10.1016/j.eswa.2023.120068
Received 17 January 2023; Received in revised form 23 March 2023; Accepted 4 April 2023
Available online 11 April 2023

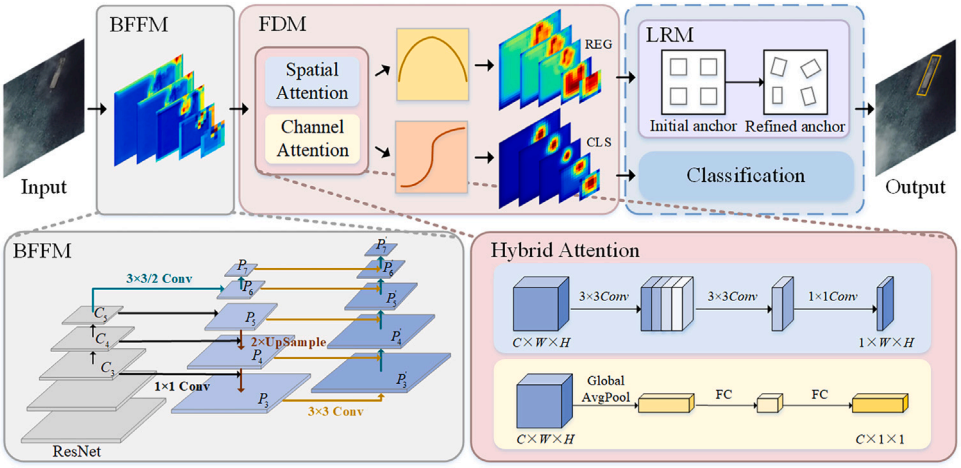


Fig. 1. The overall architecture of FDLR-Net.

While many object detection algorithms in remote sensing images have achieved relatively accurate detection results (Wang et al., 2022), there are still a number of issues that need to be further explored:

(1) The current approach does not sufficiently take into account the difficulty in achieving spatial alignment between the anchor box and the object as well as the computational effort that exists with the preset anchor boxes. The parameters of the preset anchor box have specific values, which means that the degree of alignment between the anchor box to the object region is completely dependent on the setting of the anchor box parameters. Accurate alignment requires rich prior knowledge or a large number of preset anchor boxes, which leads to complex computation and long regression process.
(2) Features extracted by the network with convolutional neural network (CNN) as backbone are rotation-invariant, which can boost the performance of classification, but are not beneficial for regressing arbitrarily oriented bounding boxes (Liao et al., 2018). The classification task and regression task sharing features will often limit detection performance.

In view of the aforementioned issues, we propose a feature decoupling and localization refinement network in this paper, aiming to generate fine regression features and achieve high alignment between anchor box and object. To enables the network to suit objects of various scales, a bidirectional feature fusion module (BFFM) is designed that generates the multi-scale feature pyramid and improve the capability of features to be represented. To provide more accurate region proposals for anchor box regression, a feature decoupling module (FDM) is designed that generate more detailed regression features and classification features respectively through fused attention with different attention functions. To achieve accurate alignment between anchor boxes and objects, a localization refinement module (LRM) is designed that adaptively learn and optimize anchor box parameters to continuously improve the alignment between anchor boxes and objects without increasing the number of anchor box parameters. Specifically, this work makes the following contributions:

(1) An innovative object detection framework is proposed that achieve detection of objects at different scales through a multi-scale pyramid structure, and accurate localization and classification through regression and classification branches.
(2) A Feature Decoupling Module (FDM) is designed to extract task-specific features, which enhances the feature representation of regression and classification branches, and can provide more accurate region information for anchor box regression, as well as richer semantic information for classification.
(3) A Localization Refinement Module (LRM) is designed to learn and optimize anchor box parameters autonomously without presetting anchor boxes with a priori knowledge, to continuously improve the ability of anchor boxes to capture object features, and to adaptively align the anchor boxes and the objects.

2. Related work

The anchor-based object detection algorithms use two kinds of anchors: horizontal anchors and rotation anchors.

Horizontal object detectors are widely used in natural scenes, and common object detection algorithms are horizontal object detectors. R-CNN (Girshick et al., 2014) first applies convolutional neural networks (CNNs) to object detection. Then Fast R-CNN (Girshick, 2015) advanced the feature extraction method of R-CNN with faster computational speed. Faster R-CNN (Ren et al., 2015) is a typical two-stage detection algorithm that classifies and regresses each proposal region using the region proposal network (RPN). Although high accuracy detection results are obtained, the detection process is time-consuming. Different from two-stage networks, YOLO (Redmon et al., 2016) transforms the object detection task into a regression task, simply using CNNs to simultaneously perform category prediction and location prediction on objects, which greatly decreases the network's computation complexity and improves the detection speed. However, YOLO has poor detection performance for densely arranged objects and small objects. SSD (Liu et al., 2016) predicts on multiple feature maps at various scales, which effectively balances the detection performance on small-scale and large-scale objects, and offers higher detection accuracy and faster detection speed. In one-stage detectors like the YOLO series, SSD and FCOS (Tian et al., 2019), positive and negative samples are extremely unbalanced, which results in their detection accuracy being lower than that of two-stage detectors. RetinaNet (Lin, Goyal, et al., 2017) uses Focal Loss to achieve a balance between positive and negative samples, while using feature pyramids to make predictions on multiple layers of feature maps, which achieves higher accuracy than Faster R-CNN.

Rotation object detectors are commonly used for text detection and remote sensing object detection. RRPN (Ma et al., 2018) and R^2CNN (Jiang et al., 2017) are both arbitrarily-oriented text detection algorithms. Both of them are built on the Faster R-CNN structure and use RPN to generate proposal regions. RRPN generate sloped proposals with angle information for text orientation, and subsequently tuned the angle information for regression of the bounding boxes to enable the proposals correspond with the text region more consistently with regard to orientation. Differently, R^2CNN generates axis-aligned bounding boxes surrounding texts in various orientations first, and then forecasts the classification score, inclined minimum area boxs and

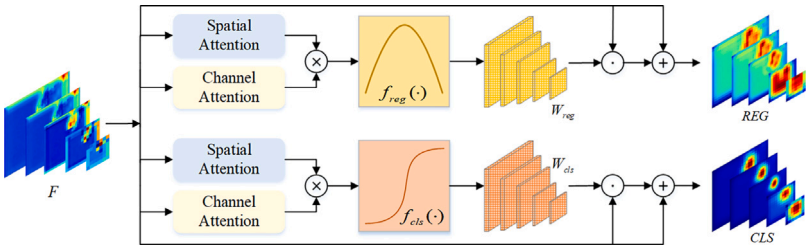


Fig. 2. The structure of FDM.

axis-aligned boxs simultaneously. SCRDet (Yang et al., 2019) merges multi-layer features with efficient anchor sampling to increase the attentiveness of small objects, while also investigates the channel attention and the supervised pixel attention for crowded object detection by and emphasizing the object feature and inhibiting the noise. RoI-Transformer (Ding et al., 2019) learns the rotation RoI through spatial transformation using horizontal anchors, and learns the transformation parameters with the supervision of angled bounding box ground-truth to fix the mismatches between the RoIs and objects for densely packed objects. R^3Det (Yang et al., 2021) develops a feature refinement module that recodes the present revised bounding box's location information into the relevant feature points via feature interpolation at the pixel level to achieve feature reconstruction and alignment.

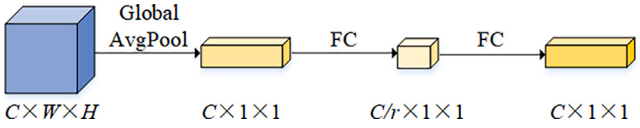


Fig. 3. Channel attention. Compression ratio r = 8.

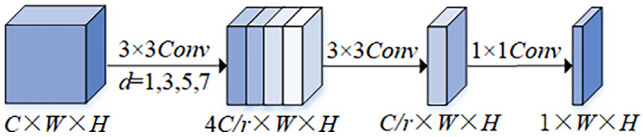


Fig. 4. Spatial attention. Compression ratio r = 8.

3. Proposed method

The overall architecture of FDLR-Net is illustrated in Fig. 1. FDLR-Net uses ResNet (He et al., 2016) as the backbone. Firstly, the bidirectional feature fusion module (BFFM) generates a multi-scale feature pyramid. Then, the features specific for classification and regression are extracted separately through the feature decoupling module (FDM). Finally, the localization refinement module (LRM) adjusts anchors in multiple levels based on the regression features. The BFFM can enhance the network's representation ability for features and the detection performance of multi-scale objects, the FDM provides the network with more accurate location information and richer semantic information to guide the network for more accurate localization and classification, the LRM can address spatial misalignment between anchor boxes and rotation objects. As a result, the detection performance can be significantly raised. Below, we give more detailed descriptions of these modules.

3.1. Bidirectional feature fusion module

In order to enables the network to suit objects of various scales and increase the network's capability for feature representation, we develop a Bidirectional Feature Fusion Module (BFFM), inspired by PANet (Liu et al., 2018). We built a bottom-up fusion path based on FPN (Lin, Dollár, et al., 2017) to reduce the layers that shallow features transmit to the top, so as to reduce the feature loss during the transmission. Considering the presence of numerous large objects in remote sensing images like large ships and airplanes, we also introduce P6 and P7 layers in the feature pyramid to predict larger scale objects. Details of BFFM are illustrated in Fig. 1.

3.2. Feature decoupling module

In most object detection networks, classification tasks and regression tasks share features. However, the feature extracted by CNN are rotation-invariant, which often boost the performance of classification, but would hinder the regression of arbitrarily oriented bounding boxes. Moreover, the spatial sensitivity contributes more to distinguishing object types, but is not robust to determine the location offset of the objects (Wu et al., 2020). In contrast, the regression task needs rotation-sensitive features and focus equally on all positions of an object. Therefore, the detection performance is limited due to sharing of features between classification and regression. In this paper, we develop a Feature Decoupling Module (FDM) to extract features specific to the regression and classification, which corresponds to enhance the location information and semantic information in the features, respectively. The structure of FDM is shown in Fig. 2. Firstly, the fusion of spatial attention (Woo et al., 2018) and channel attention (Hu et al., 2018), as well as task-specific attention functions are applied to obtain attention maps for regression and classification. Then, the task-specific attention maps decouple the input feature into the regression feature and the classification feature.

For the input feature pyramid $F \in R^{C\times W\times H}$, the classification feature and regression feature are computed as follows:

$$REG = f_{reg}(W_c \otimes W_s) \odot F + F \tag{1}$$

$$CLS = f_{cls}(W_c \otimes W_s) \odot F + F \tag{2}$$

where $\otimes$ represents matrix cross product, $\odot$ represents element-wise product. $W_c \in R^{C\times 1\times 1}$ represents the weight of different channels of F, $W_s \in R^{1\times W\times H}$ represents the weight of different pixels of F. $f_{reg}()$ and $f_{cls}()$ represent attention functions for regression and classification, respectively. $REG \in R^{C\times W\times H}$ represents the feature specific for the regression task, and $CLS \in R^{C\times W\times H}$ represents the feature specific for the classification task.

The channel attention aims to simulate the correspondence between channels. It learns each feature channel's importance and allocates different weights on each channel, so as to enhances or suppresses different channels for different tasks. Specifically, for the input feature $F \in R^{C\times W\times H}$, the channel attention model is processed as follows:

$$W_c = Sig\left(FC_2\left(FC_1\left(GAP\left(F\right)\right)\right)\right) \tag{3}$$

where GAP represents the global average pooling, FC_1 and FC_2 represent fully connected layers. Sig represents Sigmoid function, which translates the feature values to the range (0,1). Fig. 3 displays details of the channel attention.

The spatial attention aims to strengthen the feature representation in key regions, so as to enhance specific object regions of interest and weaken irrelevant background regions. Four dilated convolutions,

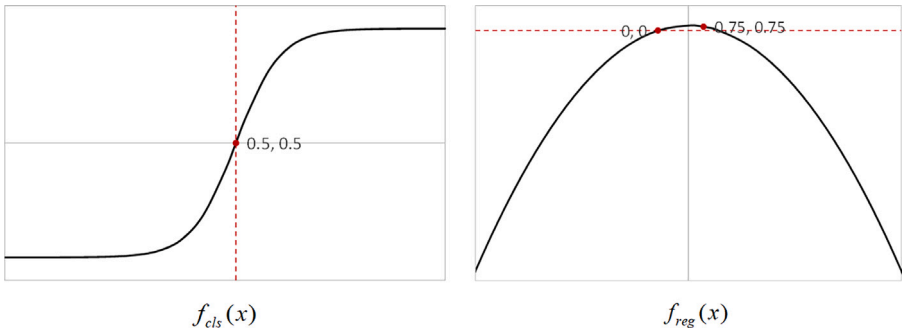


Fig. 5. The curves of activation functions.

each with a different dilation rate, are added to the spatial attention model to broaden the receptive field and reduce the spatial feature loss. Specifically, for the input feature map $F \in R^{C\times W\times H}$, the spatial attention model is processed as follows:

$$W_s = Sig(Conv^1(Conv^3(Cat(Conv_1^3, Conv_3^3, Conv_5^3, Conv_7^3)(F)))) \tag{4}$$

where $Conv^n$ represents a convolution with the $n\times n$ convolution kernel, $Conv_i^3$ represent a convolution with a convolution kernel size of 3×3 and a dilation rate of i. Cat represents the concatenate in the channel dimension. Details of the channel attention are illustrated in Fig. 4.

After the hybrid attention model, different activation functions for classification and regression are used to obtain task-specific features. In the classification task, the semantic information of the object is given more attention. Some key features in the image are sufficient to achieve accurate classification and there is no need for too many other features. Therefore, the activation function should be able to enhance the key semantic features in the image that contribute significantly to the classification task and suppress other irrelevant features. The activation functions for the classification task is designed as follows:

$$f_{cls}(x) = \frac{1}{1 + e^{-(x-0.5)}} \tag{5}$$

In the regression task, the boundary features of the object are given more attention. High response in one feature region of the image is not beneficial for accurate localization. Therefore, the activation function should be able to suppress the region with high response in the regression feature and make the response of the object feature region more evenly distributed. The activation function for the regression task is designed as follows:

$$f_{reg}(x) = 4x(1 - x) \tag{6}$$

The curves of activation functions are shown in Fig. 5. For the classification branch, feature regions with attention responses larger than 0.5 are enhanced and those with response smaller than 0.5 are suppressed. For the regression branch, feature regions with attention responses larger than 0.75 are suppressed and those with response smaller than 0.75 are smoothed.

3.3. Localization refinement module

In anchor-based object detection algorithms, anchors of certain scales, aspect ratios and angles are pre-set in order to detect objects of different scales and orientations. However, pre-setting anchors will cause the following problems:

(1) Pre-setting the anchor parameters requires the prior knowledge, and the detector performance will depend seriously on these parameter values, which is inefficient as it costs lots of time to tune the parameters for optimal detector performance.
(2) Pre-setting anchors will result in a massive number of anchors and a substantial increase in computational complexity.

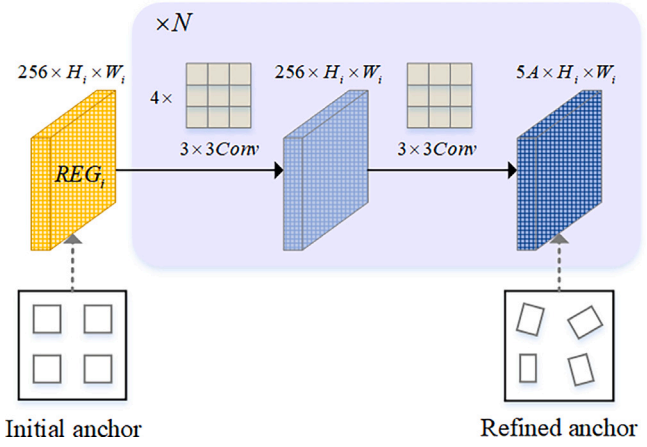


Fig. 6. The structure of LRM.

(3) Objects are highly variable in scale and have arbitrary orientation in remote sensing images, which makes it difficult to achieve alignment of the anchor and the object with a fixed anchor, and few object features are extracted by anchors, so that the detection performance is poor.

In light of the previous analysis, we design a Localization Refinement Module (LRM). LRM takes the regression feature REG as input. First, a certain number of horizontal anchors are preset at each feature point on the regression feature. Then, the anchor parameters are optimized by learning the offset between the anchors and the regression features, so as to alleviate the spatial misalignment between the anchor and the object. As illustrated in Fig. 6, for each layer REG_i of the REG, the LRM is processed as follows: first, four convolution operations with convolution kernel of 3×3 and step of 1 are conducted on the input feature map, and the number of feature channels remains constant, then a further convolution operation with convolution kernel of 3×3 and step of 1 is performed to change the number of feature channels to $5A$. A represents the anchor number at each feature point.

We use the five-parameter method to represent the arbitrary-oriented box bounding, denoted as (x, y, w, h, θ), where $\theta = 0$ for the horizontal anchor. The anchor after n levels refinement is represented as $(x_n, y_n, w_n, h_n, \theta_n)$, where $n = 0$ denotes the initial anchor. The offset learned in the $n+1$st level refinement is represented as $(t_x^{n+1}, t_y^{n+1}, t_w^{n+1}, t_h^{n+1}, t_\theta^{n+1})$, defined as follows:

$$\begin{cases} t_x^{n+1} = \dfrac{x_{n+1} - x_n}{w_n}, t_y^{n+1} = \dfrac{y_{n+1} - y_n}{h_n} \\ t_w^{n+1} = \log(\dfrac{w_{n+1}}{w_n}), t_h^{n+1} = \log(\dfrac{h_{n+1}}{h_n}) \\ t_\theta^{n+1} = \tan(\theta_{n+1} - \theta_n) \end{cases} \tag{7}$$

where (x_i, y_i), w_i, h_i and θ_i represent the center point coordinates, width, height and angle of the refined anchor at level i, respectively.

3.4. Loss function

Because of the periodicity of the angle, the bounding box described by the five-parameter will have boundary problems in regression, resulting in inaccurate regression results. Following SCRDet (Yang et al., 2019), we use IoU-Smooth $L1$ to solve this problem, which is defined as follows:

$$IoU\text{-Smooth}L_1 = -\log(IoU) \sum_{j\in\{x,y,w,h,\theta\}} \frac{L_{reg}(v'_{nj}, v_{nj})}{|L_{reg}(v'_{nj}, v_{nj})|} \tag{8}$$

where $L_{reg}(\cdot)$ represents the traditional Smooth $L1$ loss (Girshick, 2015), v'_{nj} represents the offset of the prediction, v_{nj} represents the offset of the object, and IoU indicates the intersection ratio between the predicted bounding box and the true bounding box. The regression loss can be regarded as two components, $\frac{L_{reg}(v'_j, v_j)}{|L_{reg}(v'_j, v_j)|}$ determines the gradient propagation's direction and $|-\log(IoU)|$ determines the gradient's magnitude. The rapid increase in loss is eliminated at the boundary where the loss function is almost 0.

The proposed object detection network contains classification and regression branches. We apply Focal Loss (Lin, Goyal, et al., 2017) as the classification loss, as defined below:

$$L_{cls} = \frac{1}{N}\sum_{i=1}^{N} FL(p_i, t_i) \tag{9}$$

where N represents the anchor number, $FL(\cdot)$ represents Focal Loss, t_i represents the object's label, and p_i represents the predicted probability scores of the different categories.

The regression loss consists of two parts, the localization refinement regression loss and the detection regression loss, both are given by IoU-Smooth $L1$. The localization refinement regression loss and the detection regression loss are defined as follows:

$$L_{reg_r} = \frac{1}{N}\sum_{i=1}^{N} t'_i \sum_{j\in\{x,y,w,h,\theta\}} \frac{L'_{reg}(v^r_{ij}, v_{ij})}{|L'_{reg}(v^r_{ij}, v_{ij})|} |-\log(IoU^r_g)| \tag{10}$$

$$L_{reg_p} = \frac{1}{N}\sum_{i=1}^{N} t'_i \sum_{j\in\{x,y,w,h,\theta\}} \frac{L'_{reg}(v^p_{ij}, v_{ij})}{|L'_{reg}(v^p_{ij}, v_{ij})|} |-\log(IoU^p_g)| \tag{11}$$

where N represents the number of anchors, $t'_n = 0$ represents the background and $t'_n = 1$ represents the foreground. v^r_{ij}, v_{ij} and v^p_{ij} represent the offsets of the refined anchor, the true bounding box and the predicted bounding box with respect to the preset anchor, respectively. IoU^r_g and IoU^p_g represent the intersection ratios of the refined anchor and the predicted bounding box to the true bounding box, respectively.

The total loss function of FDLR-Net is defined as follows:

$$L_{all} = L_{cls} + \lambda_1 L_{reg_r} + \lambda_2 L_{reg_p} \tag{12}$$

4. Experiments

4.1. Datasets

We have carried out experiments on DOTA dataset (Xia et al., 2018) and HRSC2016 dataset (Liu et al., 2017).

DOTA is one of the most commonly used remote sensing images dataset for object detection. DOTA contains 2806 remote sensing images, ranging in size from 800 × 800 to 4000 × 4000, with a total of 188282 instances, including 15 categories of objects. The sizes of images range from 800 × 800 to 4000 × 4000. The training set contains 1/2 of the images, the test set contains 1/3 of the images, and the validation set contains 1/6 of the images.

HRSC2016 is a high-resolution remote sensing image dataset for ship detection. HRSC2016 contains 1061 remote sensing images, ranging in size from 300 × 300 to 1500 × 900, with a total of 2976 objects, including 3 major categories and 27 minor categories. The training set contains 436 images, the test set contains 181 images, and the validation set contains 444 images.

4.2. Implementation details

The initial anchor scale is set to 2, the aspect ratio o 1, and the steps to 8, 16, 32, 64, and 128. Weight factors λ_1 and λ_2 in loss function are set to 1. The Adam optimizer is utilized to train the network, with the initial learning rate is set as 2.5×10^{-5} and decreased by 10 at each decay step. The batch size is set to 2. We pre-training the network with the warm up strategy by 3 epochs, and the learning rate is set as 1×10^{-5}. On DOTA, the training epoch is set to 30, training images are cropped to 600 × 600 with a 450-pixel overlap and random flipping is used for data augmentation. On HRSC2016, the training epoch is set to 20, training images are simply scaled to 800 × 800 without any data augmentation, and only the first category is taken, that is, all objects are treated as ships. As evaluation metrics, Average Precision (AP) and mean Average Precision (mAP) are utilized.

4.3. Ablation study

We carry out a series of ablation experiments on DOTA dataset to confirm the effect of the proposed FDM and LMR. We use ResNet50 (He et al., 2016) as the backbone and FPN (Lin, Dollár, et al., 2017) as the multi-scale feature extractor for the baseline. For fairness, experiments are performed without any data augmentation.

4.3.1. Evaluation of FDM

This subsection discusses the effect of FDM on the performance of the detector. Table 1 displays the experimental results. Compared to the baseline, the APs of all 15 categories of objects are improved after using FDM, and the mAP is improved by 1.65%. It indicates that the regression features and classification features generated by FDM through the hybrid attention mechanism and different activation functions can guide the network to perform more accurate classification and regression, thus optimizing the object detection performance. The visualization results of features before and after using FDM is shown in Fig. 7. It can be seen that in the classification feature map, the key feature regions that can characterize the object category have high responses, while other regions have low responses; in the regression feature map, the responses are relatively evenly distributed across the whole object region.

4.3.2. Evaluation of LRM

This subsection discusses the effect of LRM and different anchor generation strategies on the performance of the detector. The different experimental setups are as follows:

A: Baseline, fifteen anchors with scales of $[2^0, 2^{1/3}, 2^{2/3}]$ and aspect ratios of [1, 1/3, 3, 5, 1/5] are set.
B: Baseline + LRM, multi-scale features extracted by FPN are used as input of LRM.
C: Baseline + FDM + LRM, classification features generated by FDM are used as input of LRM.
D: Baseline + FDM + LRM, regression features generated by FDM are used as input of LRM.

Table 2 displays the experimental results. By applying LRM to the baseline, the mAP of the detector is improved by 1.64%, and when FDM and LRM are used in conjunction, the mAP is further improved by 0.69%. It proves that LRM can improve detection performance. From the results of B, C and D, it can be seen that poorer performance is achieved by using classification features than shared features for

Table 1
Ablation Study of FDM (%).

Methods	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
Baseline	82.93	72.29	40.25	61.78	66.32	69.92	73.58	83.40	80.46	81.81	59.04	56.49	62.04	63.08	53.21	67.10
+ FDM	83.56	74.18	43.30	61.89	70.32	70.27	75.58	84.49	82.32	83.10	61.90	57.28	64.18	64.24	54.53	68.74
improvement	0.63	1.89	3.05	0.11	4.00	0.35	2.00	1.09	1.86	1.29	2.86	0.79	2.14	1.16	1.32	1.64

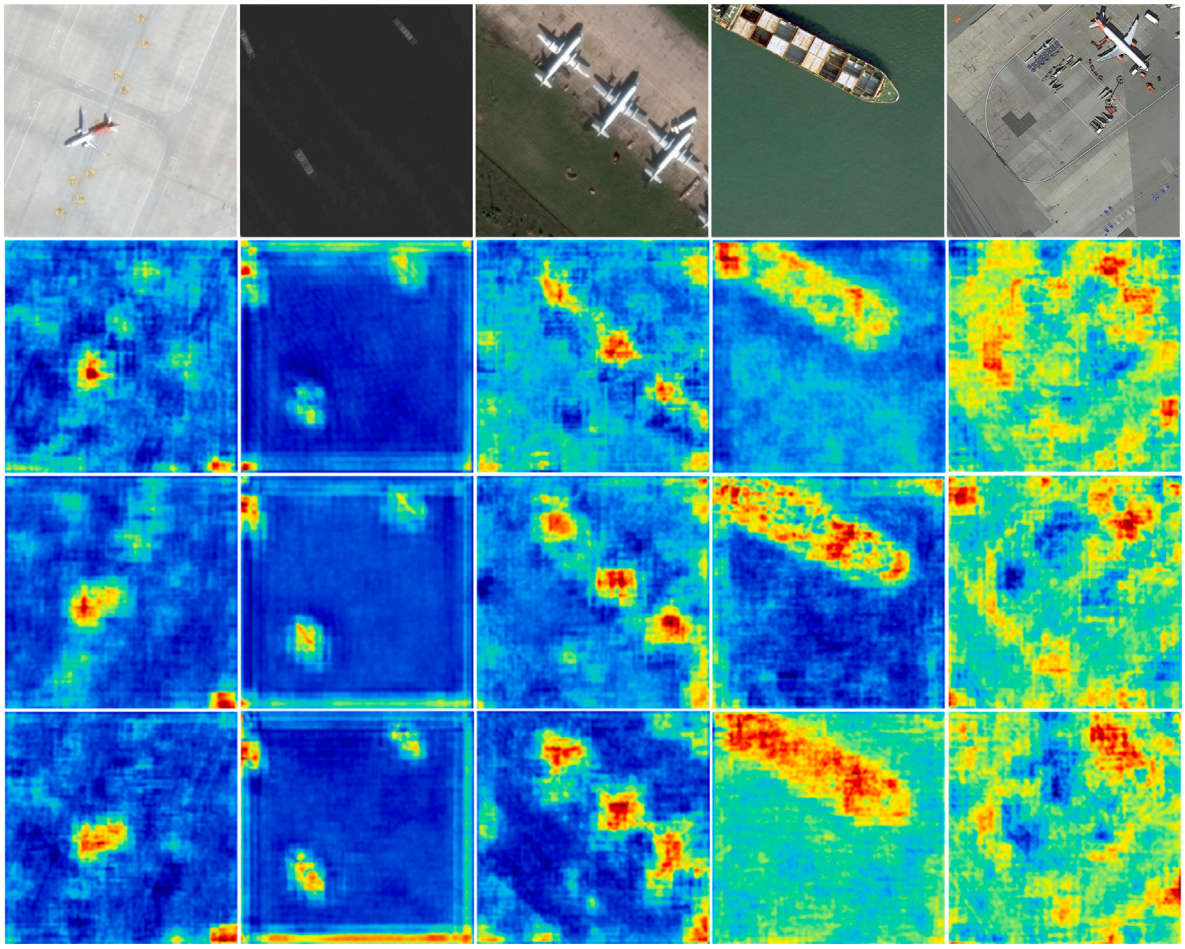


Fig. 7. The visualization of features before and after using FDM. From top to bottom are input images, shared features, classification features and regression features.

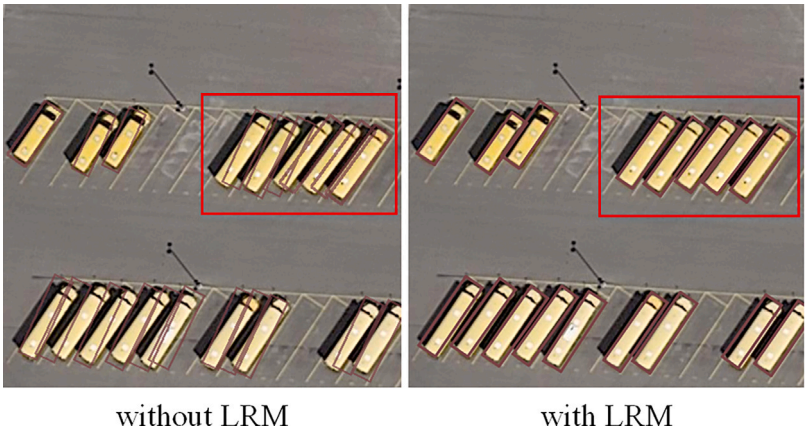


Fig. 8. The detection result visualization before and after using LRM.

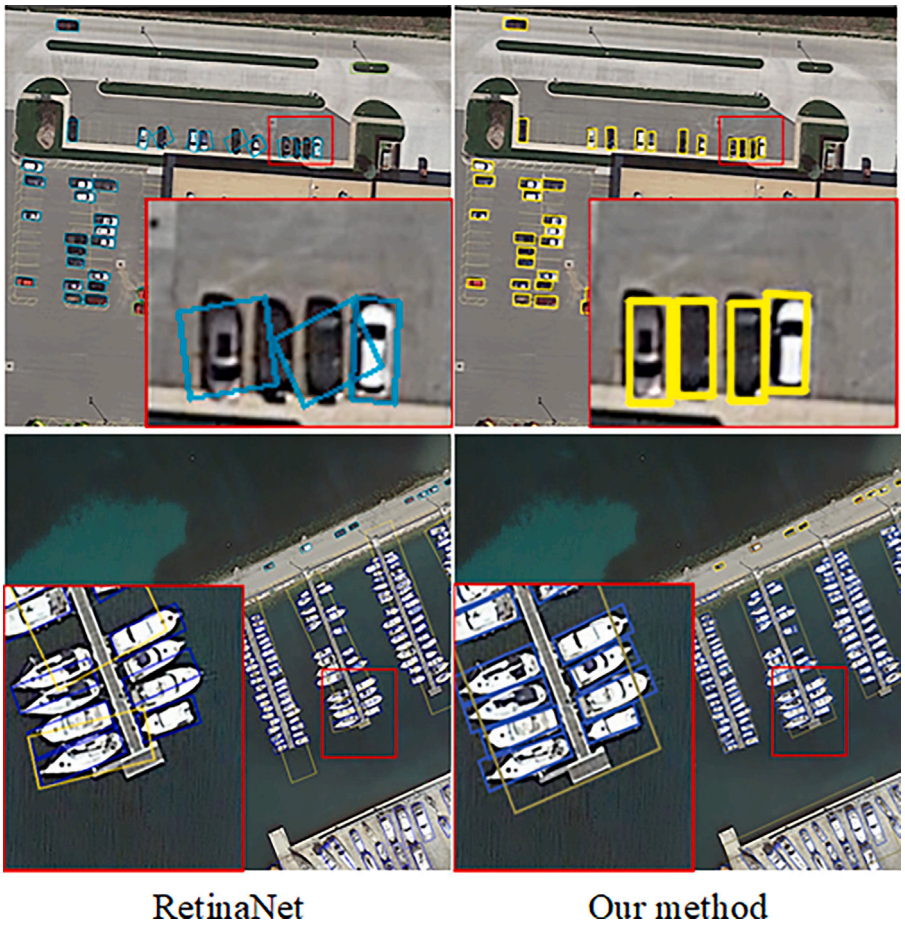


Fig. 9. Detection results visualization details of FDLR-Net in densely arranged scenes.

Table 2
Ablation Study of LRM and different anchor generation strategies.

Methods	+ FDM	+ LRM		mAP (%)
		CLS	REG	
Baseline				67.10
Baseline + LRM			✓	68.74
Baseline + FDM + LRM	✓	✓		68.33
Baseline + FDM + LRM	✓		✓	69.43

regression, the mAP is decreased by 0.41%, and better performance is achieved by using regression features instead of shared features for regression, the mAP is improved by 0.69%. It further demonstrates that the regression features generated by FDM can improve the localization accuracy of detector, and the classification features are detrimental to localization accuracy because of their high response to semantic features. Fig. 8 illustrates the detection result visualization before and after the use of LRM. With the use of LRM, the offset between predicted bounding box and object is minimized, making the localization more accurate.

4.4. Comparison with the state-of-the-art

4.4.1. Results on DOTA

We compare FDLR-Net with R^{2}CNN (Jiang et al., 2017), RRPN (Ma et al., 2018), RetinaNet (Lin, Goyal, et al., 2017), CAD-Net (Zhang et al., 2019), SCRDet (Yang et al., 2019), RoI-Transformer (Ding et al., 2019), R^{3}Det (Yang et al., 2021) and method (Xiao et al., 2022) on DOTA dataset (Xia et al., 2018). The proposed method uses ResNet152 as the backbone. Table 3 displays the experimental results. The proposed method achieves the mAP of 73.08%, which is the best among the nine algorithms, and achieves the top AP on five categories of objects, including small vehicle, large vehicle, bridge, harbor and swimming pool. Among all objects, bridges, small vehicles, large vehicles and ships have large aspect ratios, which makes it more difficult to achieve spatial alignment between the anchor and the object. Our FDLR-Net achieves excellent detection performance on these four categories of objects, which is attributed to that the proposed FDM can accurately extract features representing the object region, and then LRM guides the anchor to align with the true bounding box based on these features. At the same time, we also compare the number of parameters and training time of each method in Table 4. It can be seen that the proposed method has the least number of parameters. Compared with the method of preset anchor frame, FDLR-Net only learns five parameters($t_x, t_y, t_w, t_h, t_\theta$), which is designed lightest. In terms of training time, the proposed method is at an average level. On the whole, FDLR-Net has both good accuracy and speed. The experimental results validate the remarkable performance of FDLR-Net.

Fig. 9 illustrates the detection results visualization details of FDLR-Net for 15 categories objects on DOTA dataset. Our method obtains good detection results for different scenes, scales and directions. Fig. 10 illustrates the detection result details of RetinaNet and the proposed method for dense and small objects. From the red boxes, it can be seen that RetinaNet has both missed and false detections when detecting small vehicles and dense ships, while the proposed method has excellent detection results.

4.4.2. Results on HRSC2016

We compare FDLR-Net with RRPN (Ma et al., 2018), R^{2}CNN (Jiang et al., 2017), RoI-Transformer (Ding et al., 2019), R^{3}Det (Yang et al.,

Table 3
Detection performance on DOTA (%). The optimal detection performance of each category are bolded.

Methods	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
R^2CNN	80.94	65.75	35.34	67.44	59.92	50.91	55.81	90.67	66.92	72.39	55.06	52.23	55.14	53.35	48.22	60.67
RRPN	88.52	71.20	31.66	59.30	51.85	56.19	57.25	90.81	72.84	67.38	56.69	52.87	53.08	51.94	53.58	61.01
RetinaNet	87.93	81.64	43.69	66.69	69.29	54.77	73.26	90.74	80.22	75.54	53.89	62.91	63.90	65.93	52.29	68.72
RoI-Trans.	88.64	78.52	43.44	**75.92**	68.81	73.68	**83.59**	90.74	77.27	81.46	58.38	53.54	62.83	58.93	47.67	69.56
CAD-Net	87.80	82.40	49.40	73.50	71.10	63.50	76.70	90.90	79.20	73.30	48.40	60.90	62.00	67.00	62.20	69.90
Method (Xiao et al., 2022)	88.70	**82.63**	49.12	64.84	70.63	59.92	77.10	**91.48**	83.66	78.47	56.95	63.44	64.53	67.82	55.63	70.33
SCRDet	**89.98**	80.65	52.09	68.36	68.36	60.32	72.41	90.85	**87.94**	**86.86**	**65.05**	**66.68**	66.25	68.24	**65.21**	72.61
R^3Det	89.54	81.99	48.46	62.52	70.48	74.29	77.54	90.80	81.39	83.54	61.97	59.82	65.44	67.46	60.05	71.69
FDLR-Net	89.04	79.16	**52.10**	68.60	**72.12**	**75.08**	77.91	89.42	86.73	86.34	64.84	61.40	**66.91**	**68.65**	57.93	**73.08**

Fig. 10. Detection results visualization of FDLR-Net for 15 categories objects on DOTA dataset.

Table 4
Number of parameters and training time for the compared models.'-' indicates that there is no corresponding data.

Methods	Param	Train Time
R^2CNN	89.8M	–
RRPN	92.3M	–
RetinaNet	94.6	–
RoI-Trans.	273M	0.475s
CAD-Net	–	–
Method (Xiao et al., 2022)	101.8M	–
SCRDet	106.8M	1.18s
R^3Det	217.5M	–
FDLR-Net	74.5M	0.79s

Table 5
Detection results on HRSC2016. The optimal detection results of each category are bolded.

Methods	Backbone	Image size	mAP (%)
R^2CNN	ResNet101	800 × 800	73.7
RRPN	ResNet101	800 × 800	79.1
RoI-Trans.	ResNet101	512 × 800	86.2
Method (Xiao et al., 2022)	ResNet50	800 × 800	87.1
Method (Xiao et al., 2022)	ResNet101	800 × 800	87.7
R^3Det	ResNet101	800 × 800	88.3
FDLR-Net	ResNet50	800 × 800	88.2
FDLR-Net	ResNet101	800 × 800	88.5
FDLR-Net	ResNet152	800 × 800	**89.4**

2021) and method (Xiao et al., 2022) on HRSC2016 dataset (Xia et al., 2018). Table 5 displays the experimental results. FDLR-Net achieves the best performance on all backbones, and when ResNet152 is used as the backbone, the mAP of FDLR-Net reaches 89.4%, which is the best among all frameworks. Fig. 11 illustrates the detection results visualization of FDLR-Net with ships of different scales in different scenes, including ships covered by light clouds, ships sailing at sea and ships parking in port. The proposed method performs well in all cases.

5. Conclusion

We have proposed an end-to-end rotation detector called FDLR-Net for objects that have large scale-variation and arbitrary orientation in remote sensing images. Considering the difficulty in achieving spatial alignment between the anchor box and the object as well as the computational that exists with the preset anchor boxes, we design a localization refinement module, which realize automatic optimization of anchor box parameters and can continuously improve the alignment between anchor boxes and objects without extra anchor box parameters. For more accurate detection, a feature decoupling module is designed to generate refined regression features for LMR with more accurate region information and refined classification features with richer semantic information. Experiment results on DOTA dataset and HRSC2016 dataset have shown the excellent detection performance of our method.

Fig. 11. Detection results visualization in different scenes with ships of different scales on HRSC2016 dataset. From top to down are ships covered by light clouds, ships sailing at sea and ships parking in port.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work is funded by the National Natural Science Foundation of China (Grant No. 42101448),the Key Research and Development Projects in Hubei Province under Grants (Grant No.2021BLB149). The numerical calculations in this article have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

References

Azimi, S. M., Vig, E., Bahmanyar, R., Körner, M., & Reinartz, P. (2018). Towards multi-class object detection in unconstrained remote sensing imagery. In *Asian conference on computer vision* (pp. 150–165). Springer.

Bochkovskiy, A., Wang, C.-Y., & Liao, H.-Y. M. (2020). Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. In *European conference on computer vision* (pp. 213–229). Springer.

Chen, S., Zhao, J., Zhou, Y., Wang, H., Yao, R., Zhang, L., & Xue, Y. (2023). Info-FPN: An informative feature pyramid network for object detection in remote sensing images. *Expert Systems with Applications*, *214*, Article 119132.

Ding, J., Xue, N., Long, Y., Xia, G.-S., & Lu, Q. (2019). Learning RoI transformer for oriented object detection in aerial images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2849–2858).

Girshick, R. (2015). Fast R-CNN. In *Proceedings of the IEEE international conference on computer vision* (pp. 1440–1448).

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 580–587).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7132–7141).

Jalayer, S., Sharifi, A., Abbasi-Moghadam, D., Tariq, A., & Qin, S. (2022). Modeling and predicting land use land cover spatiotemporal changes: A case study in Chalus watershed, Iran. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *15*, 5496–5513.

Jiang, Y., Zhu, X., Wang, X., Yang, S., Li, W., Wang, H., Fu, P., & Luo, Z. (2017). R2CNN: Rotational region CNN for orientation robust scene text detection. arXiv preprint arXiv:1706.09579.

Liao, M., Zhu, Z., Shi, B., Xia, G.-s., & Bai, X. (2018). Rotation-sensitive regression for oriented scene text detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5909–5918).

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2117–2125).

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 2980–2988).

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., & Berg, A. C. (2016). SSD: Single shot multibox detector. In *European conference on computer vision* (pp. 21–37). Springer.

Liu, S., Qi, L., Qin, H., Shi, J., & Jia, J. (2018). Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8759–8768).

Liu, Z., Yuan, L., Weng, L., & Yang, Y. (2017). A high resolution optical satellite image dataset for ship recognition and some new baselines. In *International conference on pattern recognition applications and methods, vol. 2* (pp. 324–331). SciTePress.

Ma, J., Shao, W., Ye, H., Wang, L., Wang, H., Zheng, Y., & Xue, X. (2018). Arbitrary-oriented scene text detection via rotation proposals. *IEEE Transactions on Multimedia*, *20*(11), 3111–3122.

Neves, G., Ruiz, M., Fontinele, J., & Oliveira, L. (2020). Rotated object detection with forward-looking sonar in underwater applications. *Expert Systems with Applications*, *140*, Article 112870.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 779–788).

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, *28*.

Tian, Z., Shen, C., Chen, H., & He, T. (2019). Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 9627–9636).

Wahla, S. S., Kazmi, J. H., Sharifi, A., Shirazi, S. A., Tariq, A., & Joyell Smith, H. (2022). Assessing spatio-temporal mapping and monitoring of climatic variability using SPEI and RF machine learning models. *Geocarto International*, 1–20.

Wang, Y., Bashir, S. M. A., Khan, M., Ullah, Q., Wang, R., Song, Y., Guo, Z., & Niu, Y. (2022). Remote sensing image super-resolution and object detection: Benchmark and state of the art. *Expert Systems with Applications*, *197*, Article 116793.

Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S. (2018). Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision* (pp. 3–19).

Wu, Y., Chen, Y., Yuan, L., Liu, Z., Wang, L., Li, H., & Fu, Y. (2020). Rethinking classification and localization for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10186–10195).

Xia, G.-S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Datcu, M., Pelillo, M., & Zhang, L. (2018). DOTA: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3974–3983).

Xiao, J., Guo, H., Yao, Y., Zhang, S., Zhou, J., & Jiang, Z. (2022). Multi-scale object detection with the pixel attention mechanism in a complex background. *Remote Sensing*, *14*(16), 3969.

Xiao, J., Guo, H., Zhou, J., Zhao, T., Yu, Q., Chen, Y., & Wang, Z. (2023). Tiny object detection with context enhancement and feature purification. *Expert Systems with Applications*, *211*, Article 118665.

Yang, X., Yan, J., Feng, Z., & He, T. (2021). R3det: Refined single-stage detector with feature refinement for rotating object. In *Proceedings of the AAAI conference on artificial intelligence, vol. 35, no. 4* (pp. 3163–3171).

Yang, X., Yang, J., Yan, J., Zhang, Y., Zhang, T., Guo, Z., Sun, X., & Fu, K. (2019). SCRdet: Towards more robust detection for small, cluttered and rotated objects. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 8232–8241).

Zhang, Z., Guo, W., Zhu, S., & Yu, W. (2018). Toward arbitrary-oriented ship detection with rotated region proposal and discrimination networks. *IEEE Geoscience and Remote Sensing Letters*, *15*(11), 1745–1749.

Zhang, G., Lu, S., & Zhang, W. (2019). CAD-Net: A context-aware detection network for objects in remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, *57*(12), 10015–10024.