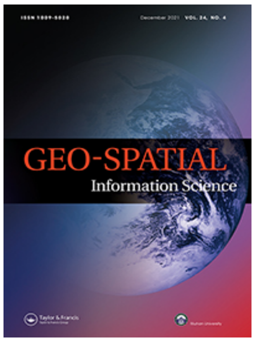

Geo-spatial Information Science

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tgsi20

An internal-external optimized convolutional neural network for arbitrary orientated object detection from optical remote sensing images

Sihang Zhang, Zhenfeng Shao, Xiao Huang, Linze Bai & Jiaming Wang

To cite this article: Sihang Zhang, Zhenfeng Shao, Xiao Huang, Linze Bai & Jiaming Wang (2021) An internal-external optimized convolutional neural network for arbitrary orientated object detection from optical remote sensing images, Geo-spatial Information Science, 24:4, 654-665, DOI: 10.1080/10095020.2021.1972772

To link to this article: https://doi.org/10.1080/10095020.2021.1972772

Published online: 27 Sep 2021.

GEO-SPATIAL INFORMATION SCIENCE
2021, VOL. 24, NO. 4, 654–665
https://doi.org/10.1080/10095020.2021.1972772

Taylor & Francis
Taylor & Francis Group

An internal-external optimized convolutional neural network for arbitrary orientated object detection from optical remote sensing images

Sihang Zhang[a], Zhenfeng Shao[a], Xiao Huang[b], Linze Bai[a] and Jiaming Wang[a]

[a]State Key Laboratory of Information Engineering in Surveying Mapping and Remote Sensing, Wuhan University, Wuhan, China;
[b]Department of Geosciences, University of Arkansas, Fayetteville, NC, USA

ABSTRACT

Due to the bird's eye view of remote sensing sensors, the orientational information of an object is a key factor that has to be considered in object detection. To obtain rotating bounding boxes, existing studies either rely on rotated anchoring schemes or adding complex rotating ROI transfer layers, leading to increased computational demand and reduced detection speeds. In this study, we propose a novel internal-external optimized convolutional neural network for arbitrary orientated object detection in optical remote sensing images. For the internal optimization, we designed an anchor-based single-shot head detector that adopts the concept of coarse-to-fine detection for two-stage object detection networks. The refined rotating anchors are generated from the coarse detection head module and fed into the refining detection head module with a link of an embedded deformable convolutional layer. For the external optimization, we propose an IOU balanced loss that addresses the regression challenges related to arbitrary orientated bounding boxes. Experimental results on the DOTA and HRSC2016 benchmark datasets show that our proposed method outperforms selected methods.

ARTICLE HISTORY
Received 26 January 2021
Accepted 21 August 2021

KEYWORDS
Arbitrary orientated object detection; optical remote sensing image; single-shot; deep learning

1. Introduction

With the advancement of earth observation technology, remote sensing images at meter-level or even sub-meter -level resolution have entered the market, providing data support for scientific research and practical applications (Yu et al. 2018; Wang et al. 2012). Object detection (Zhao et al. 2019) in optical remote sensing images (Bin and Li 2004; Han et al. 2014) has played an important role in the military as well as civilian fields. Facing a massive volume of remote sensing data with great heterogeneity, however, exaggerates the limitations of traditional machine learning methods (Felzenszwalb et al. 2009; Lee et al. 2011) given their limited feature capturing capability. In order to use remote sensing data more effectively, deep learning (LeCun, Bengio, and Hinton 2015) based object detection methods (Shao et al. 2019; Zhang et al. 2020; Shao et al. 2021) were developed and have been proved to outperform other methods.

Deep learning-based object detection methods were first developed for natural scene images. The advent of Region-based Convolutional Neural Networks (R-CNNs) (Girshick et al. 2014) fostered the development of two-stage object detection networks, e.g. Fast R-CNN (Girshick 2015) and Faster R-CNN (Ren et al. 2015). Despite the decent detection accuracy of two-stage networks, they are generally computationally intensive, thus leading to a long processing time. To solve the problem of the slow detection speed of the two-stage detection networks, one-stage object detection networks, represented by YOLOv1-3 (Redmon et al. 2016; Redmon and Farhadi 2017, 2018) and SSD (Liu et al. 2016a), were developed by slightly sacrificing detection precision. In recent years, many robust object detectors have been proposed for retaining the advantages of one-stage and two-stage detection networks while resolving their shortcomings, notably RefineDet (Zhang et al. 2018) and RetinaNet (Lin et al. 2017b). The rapid development of object detection algorithms stimulated the release of training datasets to test the effectiveness of these algorithms. Popular datasets, e.g. MS COCO (Lin et al. 2014), ImageNet (Deng et al. 2009) and PASCAL VOC (Everingham et al. 2010) contain mostly natural scene images.

Although optical remote sensing images have many similarities with natural scene images, there are some noticeable discrepancies. Natural scene images, usually taken from a side view, capture information regarding the facades of objects. Due to the influence of gravity, objects are in a horizontal state if the ground camera is placed horizontally. Optical remote sensing images (Li, Wang, and Jiang 2021), usually taken from a bird's eye view, only capture upper-facade information from objects, and these objects tend to be in an arbitrary orientated state. In addition, remote sensing images can contain a massive number of densely distributed objects

CONTACT Sihang Zhang 2019206190045@whu.edu.cn

such as ships and cars. In light of these two differences, detecting an object with a horizontal bounding box in remote sensing images presumably leads to a large overlap between adjacent objects, and many detection boxes can be filtered out during Non-Maximum Suppression (NMS) (Neubeck and Van Gool 2006) processing, thus resulting in an increased omission rate. Therefore, applying arbitrary orientated bounding boxes for object detection in remote sensing images is a necessary task (Zhou et al. 2020; Fu et al. 2020; Sun et al. 2021). To address this challenge, researchers adopted a set of predefined angles for anchors so that the prior layer in a neural network can generate anchors with different orientations (Ma et al. 2018). However, the introduction of a large number of anchors increases the computational demand, leading to long training periods and detection processes. Another solution is to generate horizontal Regions of Interest (ROI) by Region Proposal Network (RPN) (Ren et al. 2015) and rotate the horizontal ROI into an arbitrary orientated ROI. Such an approach avoids the introduction of additional computation caused by anchors but results in increased algorithm complexity. In addition, regressions are difficult to conduct on rotated bounding boxes, evidenced by sudden jumps in the loss (Yang et al. 2019).

To address these challenges, we propose a novel internal-external optimized convolutional neural network that improves the detection of arbitrarily orientated objects in optical remote sensing images. In terms of internal optimization, we designed an anchor-based head detector by adopting the concept of coarse-to-fine detection from two-stage object detection networks. The refined rotated anchors are generated from the coarse detection head module and are further fed into a refining detection head module with a link to an embedded deformable convolutional layer. For external optimization, we propose a novel IOU balanced loss function to address the challenge of executing a regression on arbitrary orientated objects. Based on these two optimizations, we propose a novel single-shot detector that combines the efficiency of one-stage detection networks and the accuracy of two-stage detection networks. In addition, the algorithm we propose avoids not only the introduction of complex ROI transforming layers but also the intensive computation of anchors. We trained and tested the proposed method on DOTA (Xia et al. 2018) and HRSC2016 (Liu et al. 2016b) datasets. Experimental results demonstrate the effectiveness and capabilities of this proposed method.

The contributions of this work are summarized as follows:

(1) We designed a coarse detection head module and a refining detection head module. The refined rotating anchors are generated from the coarse detection head module and are further fed into the refining detection head module with a link of an embedded deformable convolutional layer.
(2) We propose a novel IOU balanced loss function to address the regression challenge for arbitrary orientated objects. This new loss function can mitigate regression difficulties when rotating bounding boxes are at the angle boundary.
(3) We propose a new single-shot detector to handle the arbitrary orientated object detection in optical remote sensing images. This detector embeds the above two optimizations, achieving high detection speed as well as high detection accuracy.

The remainder of this paper is organized as follows. Section 2 includes a review of the development of object detection algorithms for optical remote sensing images and the research on arbitrary orientated object detection. Section 3 presents our proposed architecture and details the internal-external optimizations of the network. Section 4 describes the datasets, experimental settings, results, as well as sensitivity analysis. Section 5 concludes the study.

2. Related work

In recent years, the advent of Region Convolutional Neural Networks (RCNNs) has boosted object detection performance. This section reviews the development of object detection algorithms for optical remote sensing images and the research on arbitrary orientated object detection.

2.1. Object detection in optical remote sensing images

Aiming at detecting objects in optical remote sensing images, the traditional machine learning-based methods regarded the object detection task as a classification problem, incorporating feature extraction (Yuan, He, and Cai 2011) and classifier learning to obtain object detection results (Yang, Xu, and Li 2017; Bai, Zhang, and Zhou 2014). Some scholars also employed saliency detection methods in object detection tasks on remote sensing images with improved performance (Fan et al. 2016). As traditional methods cannot meet the needs of large-scale real-time remote sensing image processing, deep learning-based object detection methods in optical remote sensing images have been proposed and developed in a rapid manner. The research in (Li et al. 2017) introduced a detector with a region proposal network and local-contextual feature fusion network to handle the challenge of rotation invariance and appearance ambiguity in optical remote sensing images object detection. Since CNN does not have rotation invariant features, works from (Ding et al. 2019; Yang et al. 2019; Cheng, Zhou, and Han 2016) proposed certain solutions for the arbitrary orientated object detection in optical remote sensing images. Due to a large amount of labeling tasks, researchers began to pay attention to weakly supervised deep learning based methods for optical remote sensing image object detection. Research in (Yao et al. 2020) proposed a novel method that provides only image-level labeled samples in the training stage to complete remote sensing image object detection tasks through dynamic course learning.

2.2. Arbitrary orientated object detection

Arbitrary orientated object detection originated from text detection tasks in natural scenes as texts in natural scenes can appear with different angles. Rotation Region Proposal Networks (RRPN) (Ma et al. 2018) introduced rotating ROIs to achieve scene text detection based on the RPN architecture. RRPN predesigned a total of six rotation angles for anchors so that the network can generate anchors with different orientations. However, the introduction of additional anchors unavoidably reduces the efficiency of the algorithm. Based on horizontal anchors, ROI-Trans (Ding et al. 2019) obtained a rotating ROI using fully connected layers in the RPN stage. Different from RRPN with many anchor orientational settings, ROI-Trans learned the rotating ROI from the horizontal anchors, thus greatly reducing computations. In addition, SCRDet (Yang et al. 2019) predicted rough ROI through RPN and realized the location prediction of the orientated objects via the detection head. R2CNN (Jiang et al. 2017) proposed a new strategy to detect rotation bounding boxes by predicting the height of the bounding box as well as the coordinates of the first two vertices among the four vertices in clockwise order. RR-CNN (Liu et al. 2017) proposed an RRoI pooling layer to extract features of orientated objects.

However, these methods all belong to two-stage detection networks, limited by low detection speeds caused by their high structural complexity. Researchers have also explored the possibility of single-shot arbitrary orientated object detection networks. EAST (Zhou et al. 2017), an end-to-end training detection network, rendered a new way to define rotation objects by predicting the distances between the feature points and the four sides of the rotation box and the angle information. In light of the time-consuming nature of calculating Intersection Over Union (IOU) for orientated bounding boxes, TextBoxes++ (Liao, Shi, and Bai 2018) cascaded NMS to accelerate the IOU calculation. The IOU of the smallest bounding rectangle of all boxes was calculated, and NMS with a threshold of 0.5 was selected to reduce the number of target boxes. An NMS with a threshold of 0.2 was selected on the basis of the calculated IOU of the orientated bounding box. In more recent efforts, anchor free-orientated object detectors, such as IENet (Lin et al. 2019), were proposed to avoid the calculation of anchors. The head of IENet contains three branches, each of which handles different tasks: the classification branch handles the classification task, the regression branch handles the prediction of bounding boxes, and the rotation branch handles the prediction of orientations.

3. Methodology

In this study, we propose an internal-external optimized convolutional neural network for arbitrary orientated object detection. For the internal optimization, we propose a coarse-to-fine head detector with a deformable convolutional layer embedding between the two phases to learn the deformable features. For the external optimization, we propose an IOU balanced loss function to address the regression challenge for arbitrary orientated bounding boxes.

3.1. Overview of the proposed architecture

The proposed method is an anchor-based single-shot detector, as shown in Figure 1. This network adopts ResNet101 (He et al. 2016), a widely used deep learning backbone in various fields of image processing, as the backbone. A Feature Pyramid Network (FPN) (Lin et al. 2017a) is used to learn multiscale pyramid feature maps, denoted by $\{P_3,P_4,P_5,P_6,P_7\}$, where the subscript represents the level of the feature map. FPN takes a bottom-up and a top-down path to transfer multilayer features into integrated pyramid features through lateral connections. In order to achieve both high detection precision and high detection speed, we propose a coarse-to-fine single-shot detector. The rotating candidate anchors are refined from the horizontal anchors in the coarse detection head module and then fed into the refining detection head module. To address the arbitrary orientated bounding box regression task, we embedded a deformable convolution (Dai et al. 2017) layer that supports deformable feature learning between the two modules, and implemented the multi-task loss function as the link to achieve single-shot detection.

3.2. Internal optimization mechanism

The high detection accuracy of the two-stage detectors benefits from the coarse-to-fine anchoring strategy at the sacrifice of detection speed. Despite the high detection efficiency of one-stage detection networks, they often fall short in detection accuracy compared with two-stage networks. To mitigate the shortcomings of one-stage and two-stage networks and retain their

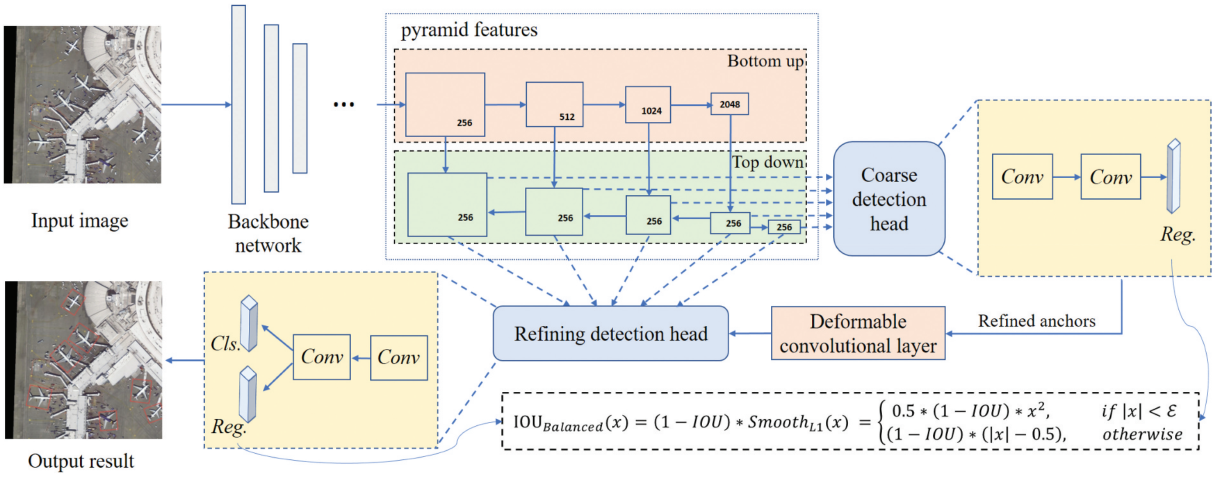
Input image
Backbone network
pyramid features
Bottom up
256
512
1024
2048
Top down
256
256
256
256
256
Coarse detection head
Conv
Conv
Reg.
Refined anchors
Deformable convolutional layer
Refining detection head
Cls.
Reg.
Conv
Conv
Output result
$\mathrm{IOU}_{Balanced}(x) = (1-IOU) * Smooth_{L1}(x) = \begin{cases} 0.5 * (1-IOU) * x^2, & if\ |x| < \mathcal{E} \\ (1-IOU) * (|x|-0.5), & otherwise \end{cases}$

Figure 1. The architecture of the proposed network contains the backbone network, the coarse detection head and the Refining detection head, where the Backbone network extracts pyramid features of input images with feature pyramid network (FPN) (Lin et al. 2017a), the coarse detection head obtains a set of refined anchors and the refining detection head aims to achieve accurate object detection result. The regression loss used in the proposed network is IOU balanced loss.

advantages, we propose a coarse-to-fine anchoring strategy in the single-shot detection network, consisting of a coarse detection head module and a refining detection head module. With the coarse detection head module, a fixed number of horizontal anchors are generated from the multiscale pyramid feature maps, followed by the implementation of a regressor to obtain a set of refined, positive rotated anchors by filtering out a large number of negative anchors. The refined rotating anchors are then passed into the refining detection head module to obtain the accurate location of arbitrary orientated bounding boxes and their categories. The coarse detection head and the refining detection head both have two convolution layers (Lawrence et al. 1997) with a kernel size of 3×3 and a stride size of 1. In order to establish a bridge between the two modules, we embed a deformable convolutional layer that learns deformable features by encoding the offsets of displacement variables. Compared with the traditional convolutional layer, the deformable convolutional layer is sampled at irregular and biased positions, which has been proved rather effective in handling the arbitrary orientated object detection in remote sensing images (Xu et al. 2017). The operation of a deformable convolutional layer can be described as:

$$y(P_0) = \sum_{P_n \in R} W(P_n) \cdot x(P_0 + P_n + t_{P_n}) \qquad (1)$$

where x is the input feature, P_0 denotes the current position, $y(P_0)$ is the output deformable feature of P_0, W is the weight coefficient matrix, t_{P_n} is the offset, and R represents the point collection of a convolution kernel grid. The term t_{P_n} can be learned by applying a convolutional layer on the same input feature map. The schematic diagram of the internal optimization mechanism is shown in Figure 2.

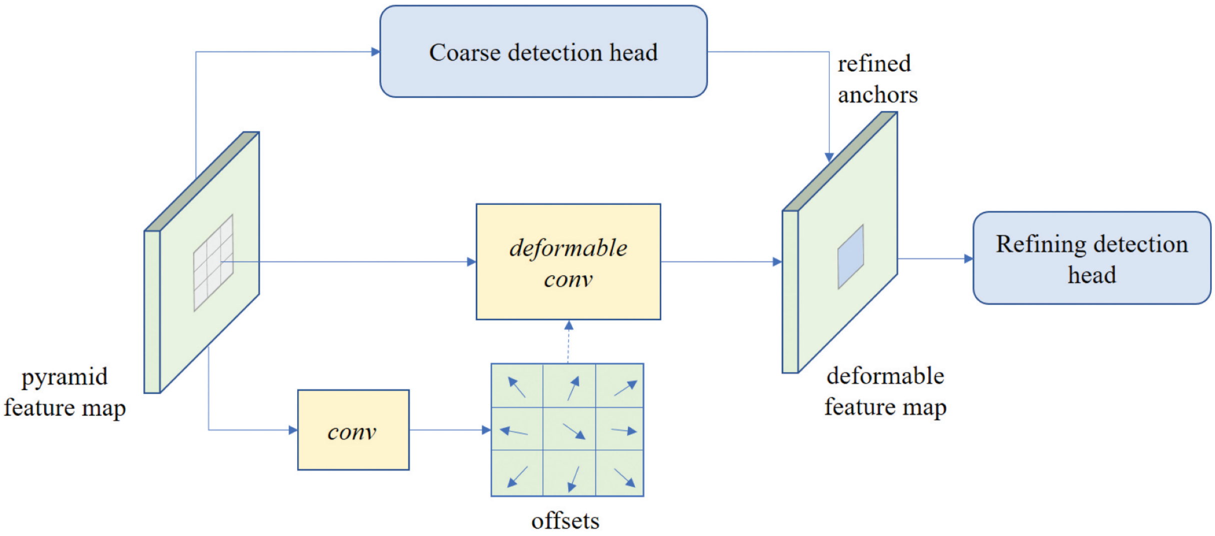

Figure 2. The schematic diagram of the internal optimization mechanism.

3.3. External optimization mechanism

Many loss functions have been proposed to deal with the bounding box regression. However, most of them are not applicable when dealing with regression for rotated bounding boxes. Due to the arbitrary directionality of the bounding boxes, sudden jumps in loss may occur at the angle threshold boundary of the prediction box, as demonstrated in Figure 3. As a result, ordinary regression loss functions may not perform well for rotated bounding boxes, and a new loss function needs to be introduced.

To address this challenge, we add a balance factor to the basis of Smooth L1 loss. When the angle of a predicted bounding box is close to that of the ground-truthing bounding box, the IOU value is large, leading to reduced loss. The proposed IOU balanced loss function is calculated as:

$$\text{IOU Balanced}(x) = (1 - IOU) * SmoothL_1(x) = \begin{cases} 0.5 * (1 - IOU) * x^2, \text{if} |x| < \varepsilon \\ (1 - IOU) * (|x| - 0.5), \text{otherwise} \end{cases} \tag{2}$$

$$IOU = \frac{intersection}{union} \tag{3}$$

where IOU denotes the intersection over union between the predicted box and the ground truth box. x represents the regression items (t_x, t_y, t_w, t_h, t_θ). εtakes 1/9 in our experiments. We define (t_x, t_y, t_w, t_h, t_θ):

$$t_x = (\cos t_\theta \cdot (x_g - x_p) + sint_\theta \cdot (y_g - y_p))/w_p$$

$$t_y = (-\sin t_\theta \cdot (x_g - x_p) + \cos t_\theta \cdot (y_g - y_p))/h_p$$

$$t_w = \log\left(\frac{w_g}{w_p}\right) \tag{4}$$

$$t_h = \log\left(\frac{h_g}{h_p}\right)$$

$$t_\theta = (\theta_g - \theta_p)/\pi$$

Our study used five parameters to represent a rotating bounding box, that is, the center point coordinates (x,y) of the rotating bounding box, the long side w and the short side h of the rotating bounding box, and the angle θ from the positive direction of x axis to the direction of w. In this work, we set $\theta \in [0, \pi)$. Thus, (x_g, y_g, w_g, h_g, θ_g) and (x_p, y_p, w_p, h_p, θ_p) denote the ground truth box and the predicted box, respectively.

The loss function of the proposed method follows the multi-task framework that consists of object regression and object classification loss. We employ the proposed IOU balanced loss for the former and focal loss (Lin et al. 2017b) for the latter. Furthermore, the loss function is the weighted summation of the coarse detection head module and the refining detection head module, defined as:

$$L = \frac{1}{N_{\text{Coa}}}\left(\sum_i Lc(c_i^{\text{Coa}}, l_i^*) + \sum_i [l_i^* \geq 1] Lr(x_i^{\text{Coa}}, g_i^*)\right) + \frac{\lambda}{N_{\text{Ref}}}\left(\sum_i Lc(c_i^{\text{Ref}}, l_i^*) + \sum_i [l_i^* \geq 1] Lr(x_i^{\text{Ref}}, g_i^*)\right) \tag{5}$$

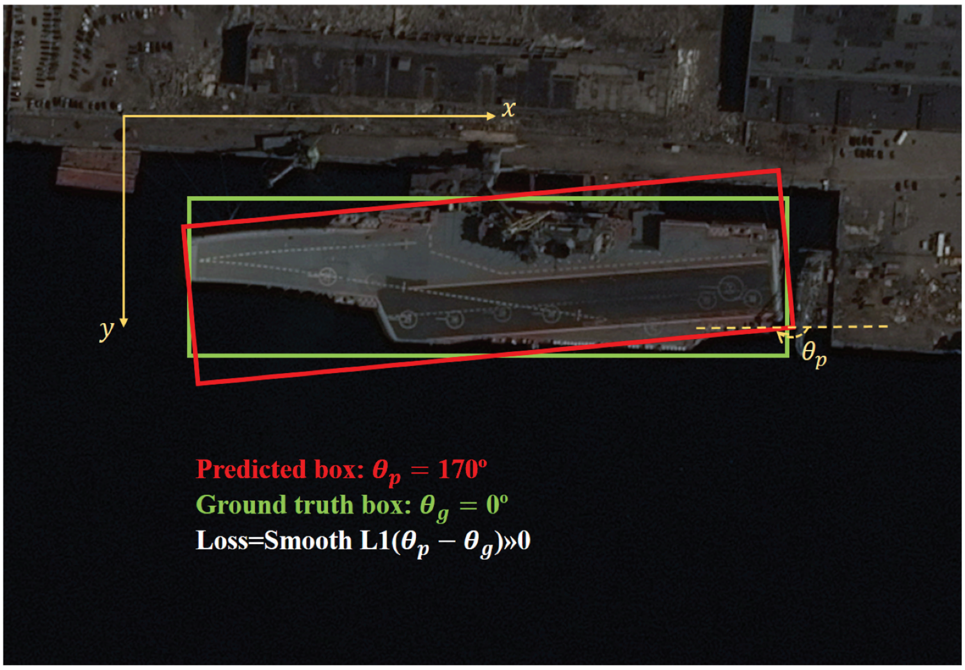
x
y
θ_p
Predicted box: θ_p = 170°
Ground truth box: θ_g = 0°
Loss=Smooth L1($\theta_p - \theta_g$)»0

Figure 3. The schematic diagram of arbitrary orientated bounding box regression using smooth L1 loss. The x axis and the y axis point to the positive direction of the screen coordinates. The term θp denotes the angle from the positive direction of x axis to the direction of the long side of the predicted box where $\theta_P \in [0, \pi)$.

where N_{Coa}, c_i^{Coa} and x_i^{Coa} denote the numbers of the positive samples, the predicted category, and the predicted bounding box location in the coarse detection head module, respectively. N_{Ref} indicates the numbers of the positive samples, c_i^{Ref} is the predicted category, and x_i^{Ref} is the predicted bounding box location in the refining detection head module. The terms Lc and Lr refer to the classification loss function and the regression loss function, respectively, while l_i^* and g_i^* represent the category and location of the ground truth box. The value λ is a balancing parameter. The Iverson bracket indicator function $[l_i^* \geq 1]$ outputs 1 when the condition (i.e.$l_i^* \geq 1$) is true (the anchor is not the negative) and 0 otherwise.

4. Experiments and analysis

4.1. Datasets

To prove the effectiveness of the proposed method, we implemented our proposed method and comparative methods on two public benchmark datasets (i.e. DOTA [Xia et al. 2018] and HRSC2016 [Liu et al. 2016b]) for object detection in high-resolution remote sensing images. Their descriptions are shown in Table 1.

4.1.1. DOTA

DOTA (Xia et al. 2018) is an aerial image dataset for object detection, which contains 2,806 aerial images with unfixed sizes from 800 × 800 to 4000 × 4000. DOTA contains 15 categories: Plane, Baseball Diamond (BD), Bridge, Ground Track Field (GTF), Small Vehicle (SV), Large Vehicle (LV), Ship, Tennis Court (TC), Basketball Court (BC), Storage Tank (ST), Soccer Ball Field (SBF), Round About (RA), Harbor, Swimming Pool (SP), and Helicopter (HC). Given the large size of the original DOTA images, which might cause memory and efficiency issues, we resized the original images at three scales (0.5, 1.0 and 1.5) and then cropped them into a series of 1024 × 1024 patches at a stride of 512.

4.1.2. HRSC2016

HRSC2016 (Liu et al. 2016b) is a dataset that contains high-resolution remote sensing images, specifically for ship detection. HRSC2016 contains 1061 images collected from google earth with 2976 instances. The objects are annotated using its center point coordinates x and y, width w and height h of the bounding box, and the angle θ between width and x-axis.

4.2. Implementation details

The proposed algorithm was implemented using PyTorch (Paszke et al. 2019) on two TITAN Xp GPUs, each with 11 G memories. We adopt ResNet101 (He et al. 2016) as the backbone network and extract pyramid features from P_3 to P_7. Stochastic Gradient Descent (SGD) is adopted to train the network with a batch size of 2 for 12 epochs on the DOTA dataset and 36 epochs on HRSC2016, with an initial learning rate of 0.0025 (divided by 10 at each decay step). The momentum and weight decay are set to 0.9 and 0.0001, respectively. In the inferencing stage, we use Non-Maximum Suppression (NMS) for post-processing.

4.3. Experimental results

Mean Average-Precision (mAP) was adopted to evaluate the performance of the object detectors. We first compared our proposed method with several existing arbitrary orientated object detectors on the DOTA data (Table 2). The results in Table 2 show that our

Table 1. Dataset description.

Data set	Images	Categories	Size	Crop size	Training images	Instances
DOTA	2806	15	800 × 800 ~ 4000 × 4000	1024 × 1024	21,046	188,282
HRSC2016	1061	1	300 × 300 ~ 1500 × 900	-	617	2976

Table 2. Detection result on DOTA.(%).

Method	RetinaNet-R	IENet	FR-O	R-DFPN	R2CNN	RRPN	ICN	ROI-Trans	Ours
Plane	88.92	57.14	79.09	80.92	80.94	88.52	81.40	88.64	88.67
BD	67.67	80.20	69.12	65.82	65.67	71.20	74.30	78.52	79.51
Bridge	33.55	64.54	17.17	33.77	35.34	31.66	47.70	43.44	52.44
GTF	56.83	39.82	63.49	58.94	67.44	59.30	70.30	75.92	70.01
SV	66.11	32.07	34.20	55.77	59.92	51.85	64.90	68.81	75.01
LV	73.28	49.71	37.16	50.94	50.91	56.19	67.80	73.68	74.00
Ship	75.24	65.01	36.20	54.78	55.81	57.25	70.00	83.59	84.03
TC	90.87	52.58	89.19	90.33	90.67	90.81	90.80	90.74	90.85
BC	73.95	81.45	69.60	66.34	66.92	72.84	79.10	77.27	80.49
ST	75.07	44.66	58.96	68.66	72.39	67.38	78.20	81.46	85.35
SBF	43.77	78.51	49.40	48.73	55.06	56.69	53.60	58.39	55.64
RA	56.72	46.54	52.52	51.76	52.23	52.84	62.90	53.54	66.08
Harbor	51.05	56.73	46.69	55.10	55.14	53.08	67.00	62.83	66.17
SP	55.86	64.40	44.80	51.32	53.35	51.94	64.20	58.93	66.25
HC	21.46	64.24	46.30	35.88	48.22	53.58	50.20	47.067	51.02
mAP	62.02	57.14	52.93	57.94	60.67	61.01	68.20	69.56	72.37

method outperformed the selected one-stage detectors based on RetinaNet-R (Lin et al. 2017b) or IENet (Lin et al. 2019) and two-stage detectors based on FR-O (Xia et al. 2018), R-DFPN (Yang et al. 2018), R2CNN (Jiang et al. 2017), RRPN (Ma et al. 2018), ICN (Azimi et al. 2018), and ROI-Trans (Ding et al. 2019). Figure 4 presents the visualization of our detection results on the DOTA dataset. Table 3 presents a performance evaluation on the HRSC2016 data comparing the proposed method with other arbitrary oriented object

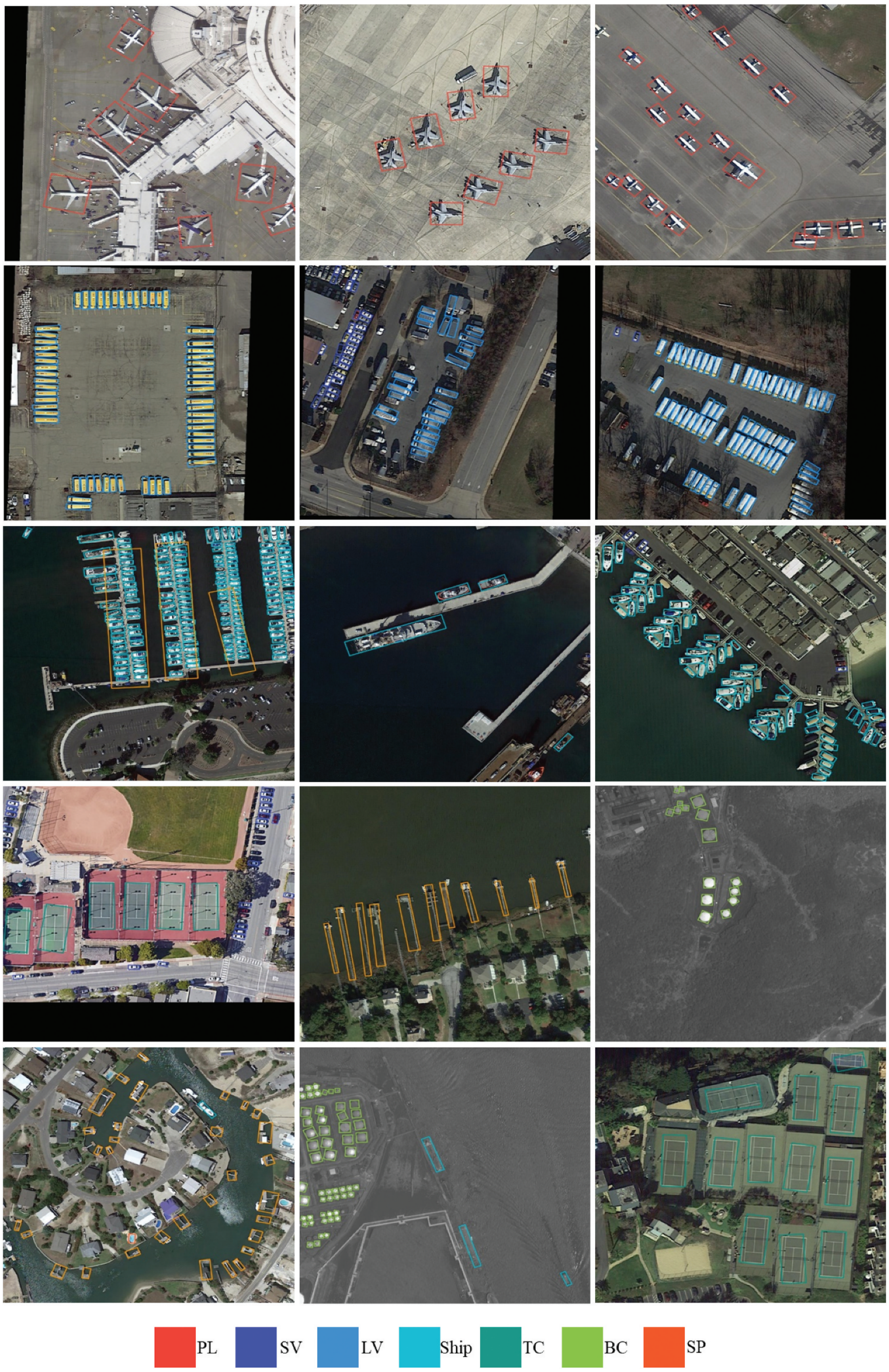

Figure 4. Selected examples of detection results from the DOTA test set.

Table 3. Detection result on HRSC2016 (%).

Method	Fast-RCNN +SRBBS	BL2 (Liu et al. 2017)	R2CNN	IENet	RRD (Liao et al. 2018)	ROI-Trans	Ours
mAP	55.70	69.60	73.07	75.01	84.30	86.20	87.26

detectors, revealing the superiority of the proposed method over selected baselines. Visual validation of our detection results on HRSC2016 is given in Figure 5.

To illustrate the effectiveness of our proposed method, we conducted speed-accuracy trade-off experiments on the DOTA among FR-O, ROI-Trans, and our proposed method (on 1024 × 1024 images). As shown in Table 4, the results indicate that our proposed method achieved the highest mAP with a detection speed of 0.137s per image, faster than FR-O and ROI-Trans. The inference time of our proposed method on the HRSC2016 dataset was 0.072s per image.

Table 4. Speed-accuracy trade-off comparison with selected methods on DOTA dataset.

Method	Image size	mAP	Test speed(s)
FR-O	1024 × 1024	52.93	0.141
ROI-Trans	1024 × 1024	69.56	0.170
Ours	1024 × 1024	72.37	0.137

4.4. Ablation studies

We conducted ablation experiments on HRSC2016 to evaluate the effectiveness of proposed optimization mechanisms. The results reveal an increase in mAP by 1.10% with the deformable convolutional layer and an increase in mAP by 8.11% with proposed IOU balanced loss (Table 5). Moreover, the mAP of the

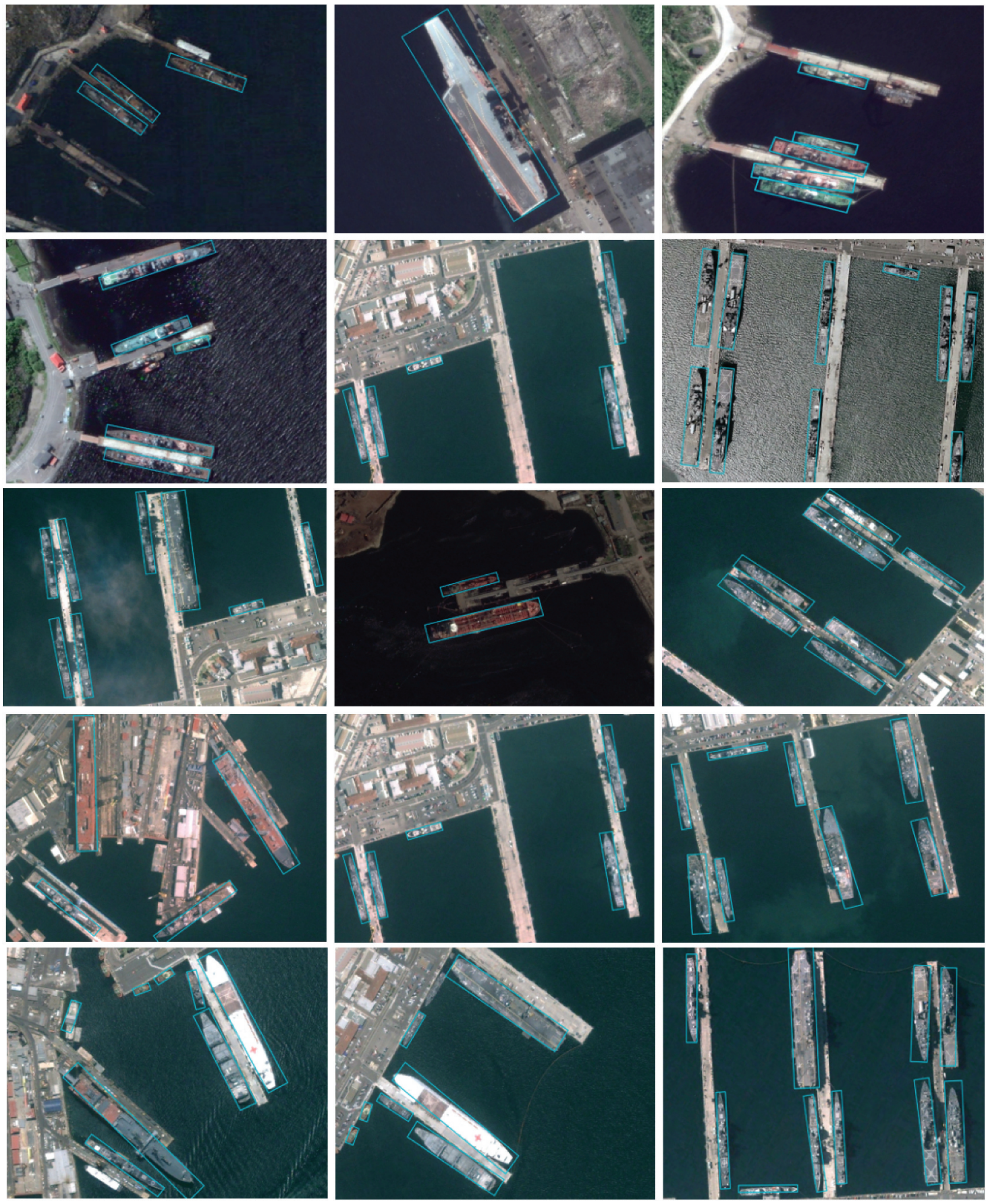

Figure 5. Selected examples of detection results on the HRSC2016 dataset.

Table 5. Ablation study on HRSC2016 dataset.

Method	Deformable conv layer	IOU balanced loss	mAP
Baseline + Smooth L1 loss	×	×	78.00
Baseline + Deformable conv layer+ Smooth L1 loss	√	×	79.10
Baseline + IOU balanced loss	×	√	86.11
Baseline + Deformable conv layer + IOU balanced loss	√	√	87.26

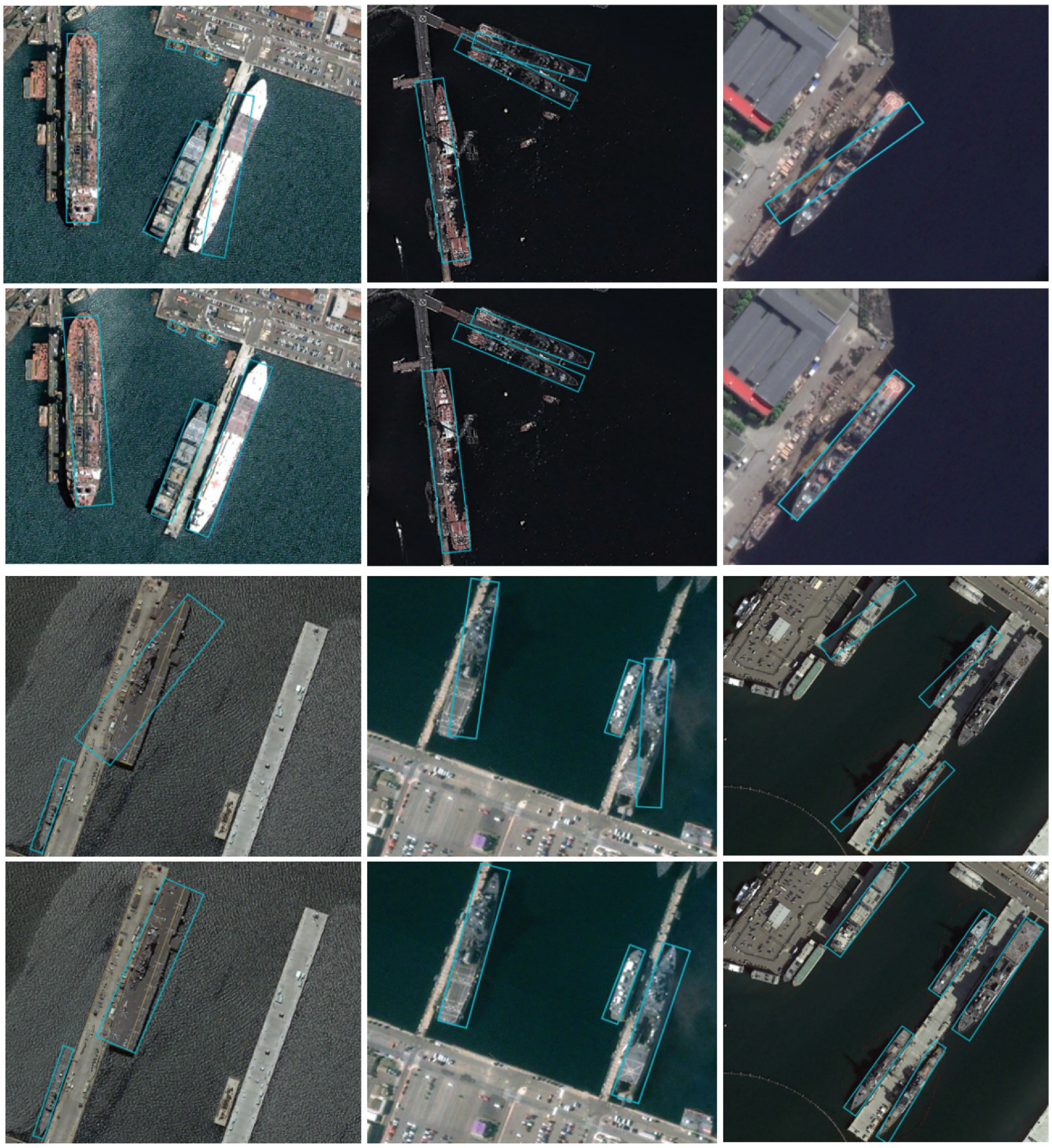

Figure 6. Comparison of detection results on the HRSC2016 dataset with IOU balanced loss (even rows) and with smooth L1 loss (odd rows).

proposed method was 9.26% higher than that of baselines, indicating the effectiveness of our proposed method. Figure 6 shows a comparison of detection results with proposed IOU balance loss and with Smooth L1 loss as the loss function, respectively.

5. Conclusions

In this paper, we propose a novel internal-external optimized convolutional neural network for arbitrary orientated object detection in optical remote sensing images. For the internal optimization, we design an anchor-based head detector that adopts the coarse-to-fine detection strategy in two-stage object detection networks. The refined rotating anchors are generated from the coarse detection head module and further fed into the refining detection head module with a link of an embedded deformable convolutional layer. For the external optimization, we propose an IOU balanced loss to address the regression challenge from arbitrary orientated bounding boxes. Integrating these two optimization mechanisms, we designed a novel single-shot detector that can handle the arbitrary orientated object detection in optical remote sensing images. Experimental results on DOTA and HRSC2016 datasets show that our proposed method outperformed other selected methods.

Disclosure statement

There are no conflicts of interest to disclose.

Data availability statement

The DOTA dataset that support the findings of this study is openly available in *https://captain-whu.github.io/DOTA/dataset.html*, and the HRSC2016 dataset that support the findings of this study is available from the corresponding author, upon reasonable request.

Funding

This work is supported by the National Natural Science Foundation of China [grant numbers 41890820, 41771452, 41771454, and 41901340]

Notes on contributors

Sihang Zhang is currently pursuing the master's degree in Wuhan University. Her research interests are remote sensing image processing and computer vision.

Zhenfeng Shao is a Professor with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University. He received the PhD degree from Wuhan University. His research interests are urban remote sensing and computer vision.

Xiao Huang received his PhD degree in Geography from the University of South Carolina in 2020. He is currently an Assistant Professor in the Department of Geosciences at the University of Arkansas with his expertise in GeoAI, deep learning, big data, remote sensing, and social sensing.

Linze Bai is currently working toward the master's degree in Wuhan University. His research interests are remote sensing image processing and computer vision.

Jiaming Wang is currently working toward the PhD degree in Wuhan University. His research field includes image/video processing and computer vision.

ORCID

Sihang Zhang http://orcid.org/0000-0002-4705-2882
Zhenfeng Shao http://orcid.org/0000-0003-4587-6826
Xiao Huang http://orcid.org/0000-0002-4323-382X
Linze Bai http://orcid.org/0000-0002-1384-5630
Jiaming Wang http://orcid.org/0000-0001-8144-5842

References

Azimi, S. M., E. Vig, R. Bahmanyar, M. Körner, and P. Reinartz. 2018. "Towards Multi-class Object Detection in Unconstrained Remote Sensing Imagery." In *Asian Conference on Computer Vision*, Perth, Australia, December 2-6.

Bai, X., H. Zhang, and J. Zhou. 2014. "VHR Object Detection Based on Structural Feature Extraction and Query Expansion." *IEEE Transactions on Geoscience and Remote Sensing* 52 (10): 6508–6520. doi:10.1109/TGRS.2013.2296782.

Bin, L., and P. Li. 2004. "Object Extraction Based on Evolutionary Morphological Processing." *Geo-spatial Information Science* 7 (3): 193–197. doi:10.1007/BF02826290.

Cheng, G., P. Zhou, and J. Han. 2016. "Learning Rotation-Invariant Convolutional Neural Networks for Object Detection in VHR Optical Remote Sensing Images." *IEEE Transactions on Geoscience and Remote Sensing* 54 (12): 7405–7415. doi:10.1109/TGRS.2016.2601622.

Dai, J., H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei. 2017. "Deformable Convolutional Networks." In *Proceedings of the IEEE international conference on computer vision*. Venice, Italy. October 22–29.

Deng, J., W. Dong, R. Socher, L. J. Li, K. Li, and F. F. Li. 2009. "Imagenet: A Large-scale Hierarchical Image Database." In *Proceedings of the IEEE conference on computer vision and pattern recognition*. Florida, America. June 20–25.

Ding, J., N. Xue, Y. Long, G. S. Xia, and Q. Lu. 2019. "Learning RoI Transformer for Oriented Object Detection in Aerial Images." In *Proceedings of the IEEE conference on computer vision and pattern recognition*. Long Beach, CA, America. June 16–20.

Everingham, M., L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. 2010. "The Pascal Visual Object Classes (VOC) Challenge." *International Journal of Computer Vision* 88 (2): 303–338. doi:10.1007/s11263-009-0275-4.

Fan, X. X., W. B. Xu, M. A. Yang, and X. S. Fan. 2016. "Target Detection in Remote Sensing Image Based on Visual Attention Mechanism." *Radio Engineering* 46 (1): 57–60.

Felzenszwalb, P. F., R. B. Girshick, D. McAllester, and D. Ramanan. 2009. "Object Detection with Discriminatively Trained Part-based Models." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32 (9): 1627–1645. doi:10.1109/TPAMI.2009.167.

Fu, K., Z. Chang, Y. Zhang, and X. Sun. 2020. "Point-based Estimator for Arbitrary-oriented Object Detection in Aerial Images." *IEEE Transactions on Geoscience and Remote Sensing* 99: 1–18.

Girshick, R. 2015. "Fast R-cnn." In *Proceedings of the IEEE international conference on computer vision*. Santiago, Chile. December 11–18.

Girshick, R., J. Donahue, T. Darrell, and J. Malik. 2014. "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation." In *Proceedings of the IEEE conference on computer vision and pattern recognition*. Columbus, Ohio, America. June 23–28.

Han, J., D. Zhang, G. Cheng, L. Guo, and J. Ren. 2014. "Object Detection in Optical Remote Sensing Images Based on Weakly Supervised Learning and High-level Feature Learning." *IEEE Transactions on Geoscience and Remote Sensing* 53 (6): 3325–3337. doi:10.1109/TGRS.2014.2374218.

He, K., X. Zhang, S. Ren, and J. Sun. 2016. "Deep Residual Learning for Image Recognition." In *Proceedings of the IEEE conference on computer vision and pattern recognition*. Las Vegas, America, June 26–July 1.

Jiang, Y., X. Zhu, X. Wang, S. Yang, W. Li, H. Wang, P. Fu, et al. 2017. "R2cnn: Rotational Region CNN for Orientation Robust Scene Text Detection." *arXiv preprint arXiv:1706.09579*.

Lawrence, S., C. L. Giles, A. C. Tsoi, and A. D. Back. 1997. "Face Recognition: A Convolutional Neural-network Approach." *IEEE Transactions on Neural Networks* 8 (1): 98–113. doi:10.1109/72.554195.

LeCun, Y., Y. Bengio, and G. Hinton. 2015. "Deep Learning." *Nature* 521 (7553): 436–444. doi:10.1038/nature14539.

Lee, J. J., P. H. Lee, S. W. Lee, A. Yuille, and C. Koch. 2011. "Adaboost for Text Detection in Natural Scene." In *International Conference on Document Analysis and Recognition*. Beijing, China, September 18–21.

Li, D., M. Wang, and J. Jiang. 2021. "China's High-resolution Optical Remote Sensing Satellites and Their Mapping Applications." *Geo-spatial Information Science* 24 (1): 85–94. doi:10.1080/10095020.2020.1838957.

Li, K., G. Cheng, S. Bu, and X. You. 2017. "Rotation-Insensitive and Context-Augmented Object Detection in Remote Sensing Images." *IEEE Transactions on Geoscience and Remote Sensing* 56 (4): 2337–2348. doi:10.1109/TGRS.2017.2778300.

Liao, M., B. Shi, and X. Bai. 2018. "Textboxes++: A Single-shot Oriented Scene Text Detector." *IEEE Transactions on Image Processing* 27 (8): 3676–3690. doi:10.1109/TIP.2018.2825107.

Liao, M., Z. Zhu, B. Shi, G. S. Xia, and X. Bai. 2018. "Rotation-sensitive Regression for Oriented Scene Text Detection." In *Proceedings of the IEEE conference on computer vision and pattern recognition*. Salt Lake City, Utah, America. June 18–23.

Lin, T. Y., P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. 2017a. "Feature Pyramid Networks for Object Detection." In *Proceedings of the IEEE conference on computer vision and pattern recognition*. Honolulu, Hawaii, America. July 21–26.

Lin, T. Y., P. Goyal, R. Girshick, K. He, and P. Dollár. 2017b. "Focal Loss for Dense Object Detection." *IEEE Transactions on Pattern Analysis & Machine Intelligence* 99: 2999–3007.

Lin, T. Y., M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, et al. 2014. "Microsoft Coco: Common Objects in Context." In *European conference on computer vision*. Zurich, Switzerland. September 6–12.

Lin, Y., P. Feng, J. Guan, W. Wang, and J. Chambers. 2019. "Ienet: Interacting Embranchment One Stage Anchor Free Detector for Orientation Aerial Object Detection." *arXiv preprint arXiv:1912.00969*.

Liu, W., D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg. 2016a. "Ssd: Single Shot Multibox Detector." In *European conference on computer vision*. Amsterdam, The Netherlands. October 11–14.

Liu, Z., J. Hu, L. Weng, and Y. Yang. 2017. "Rotated Region Based CNN for Ship Detection." In *2017 IEEE International Conference on Image Processing (ICIP)*. Beijing, China. September 17–20.

Liu, Z., H. Wang, L. Weng, and Y. Yang. 2016b. "Ship Rotated Bounding Box Space for Ship Extraction from High-Resolution Optical Satellite Images with Complex Backgrounds." *IEEE Geoscience and Remote Sensing Letters* 13 (8): 1074–1078. doi:10.1109/LGRS.2016.2565705.

Ma, J., W. Shao, H. Ye, L. Wang, H. Wang, Y. Zheng, and X. Xue. 2018. "Arbitrary-oriented Scene Text Detection via Rotation Proposals." *IEEE Transactions on Multimedia* 20 (11): 3111–3122. doi:10.1109/TMM.2018.2818020.

Neubeck, A, and L. Van Gool. 2006. "Efficient Non-maximum Suppression." In *18th International Conference on Pattern Recognition (ICPR'06)*. Hong Kong, China. August 20–24.

Paszke, A., S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, et al. 2019. "Pytorch: An Imperative Syle, High-performance Deep Learning Library." *Advances in Neural Information Processing Systems* 32: 8026–8037.

Redmon, J., S. Divvala, R. Girshick, and A. Farhadi. 2016. "You Only Look Once: Unified, Real-time Object Detection." In *Proceedings of the IEEE conference on computer vision and pattern recognition*. Las Vegas, America. June 27–30.

Redmon, J., and A. Farhadi. 2017. "YOLO9000: Better, Faster, Stronger." In *Proceedings of the IEEE Conference on computer vision and pattern recognition*. Honolulu, Hawaii, America. July 21–26.

Redmon, J., and A. Farhadi. 2018. "Yolov3: An Incremental Improvement." *arXiv preprint arXiv:1804.02767*.

Ren, S., K. He, R. Girshick, and J. Sun. 2015. "Faster R-cnn: Towards Real-time Object Detection with Region Proposal Networks"." *Advances in Neural Information Processing Systems* 28: 91–99.

Shao, Z., G. Cheng, J. Ma, Z. Wang, J. Wang, and D. Li. 2021. "Real-time and Accurate UAV Pedestrian Detection for Social Distancing Monitoring in COVID-19 Pandemic." *IEEE Transactions on Multimedia* 1. doi:10.1109/TMM.2021.3075566.

Shao, Z., L. Wang, Z. Wang, W. Du, and W. Wu. 2019. "Saliency-Aware Convolution Neural Network for Ship Detection in Surveillance Video." *IEEE Transactions on Circuits and Systems for Video Technology* 30 (3): 781–794. doi:10.1109/TCSVT.2019.2897980.

Sun, F., H. Li, Z. Liu, X. Li, and Z. Wu. 2021. "Arbitrary-angle Bounding Box Based Location for Object Detection in Remote Sensing Image." *European Journal of Remote Sensing* 54 (1): 102–116. doi:10.1080/22797254.2021.1880975.

Wang, S., C. Liu, S. Wu, Q. Nie, Y. Wang, S. Zeng, and H. Zhu. 2012. "Automatic Extraction of Foreground Objects from Mars Images." *Geo-spatial Information Science* 15 (1): 17–25. doi:10.1080/10095020.2012.708147.

Xia, G. S., X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, et al. 2018. "DOTA: A Large-scale Dataset for Object Detection in Aerial Images." In *Proceedings of the IEEE conference on computer vision and pattern recognition*. Salt Lake City, Utah, America. June 18–23.

Xu, Z., X. Xu, L. Wang, R. Yang, and F. Pu. 2017. "Deformable ConvNet with Aspect Ratio Constrained NMS for Object Detection in Remote Sensing Imagery." *Remote Sensing* 9 (12): 1312. doi:10.3390/rs9121312.

Yang, F., Q. Xu, and B. Li. 2017. "Ship Detection from Optical Satellite Images Based on Saliency Segmentation and Structure-LBP Feature." *IEEE Geoscience and Remote Sensing Letters* 14 (5): 602–606. doi:10.1109/LGRS.2017.2664118.

Yang, X., H. Sun, K. Fu, J. Yang, X. Sun, M. Yan, and Z. Guo. 2018. "Automatic Ship Detection in Remote Sensing Images from Google Earth of Complex Scenes Based on Multiscale Rotation Dense Feature Pyramid Networks." *Remote Sensing* 10 (1): 132. doi:10.3390/rs10010132.

Yang, X., J. Yang, J. Yan, Y. Zhang, T. Zhang, Z. Guo, and K. Fu. 2019. "Scrdet: Towards More Robust Detection for Small, Cluttered and Rotated Objects." In *Proceedings of the IEEE International Conference on Computer Vision*. Seoul, Korea. October 27–November 2.

Yao, X., X. Feng, J. Han, G. Cheng, and L. Guo. 2020. "Automatic Weakly Supervised Object Detection from High Spatial Resolution Remote Sensing Images via Dynamic Curriculum Learning." *IEEE Transactions on Geoscience and Remote Sensing* 59 (1): 675–685. doi:10.1109/TGRS.2020.2991407.

Yu, H., J. Wang, Y. Bai, W. Yang, and G. S. Xia. 2018. "Analysis of Large-scale UAV Images Using A Multi-scale Hierarchical Representation." *Geo-spatial Information Science* 21 (1): 33–44. doi:10.1080/10095020.2017.1418263.

Yuan, Z., Y. He, and F. Cai. 2011. "Fast Algorithm for Maneuvering Target Detection in SAR Imagery Based on Gridding and Fusion of Texture Features." *Geo-spatial Information Science* 14 (3): 169–176. doi:10.1007/s11806-011-0536-6.

Zhang, R., Z. Shao, X. Huang, J. Wang, and D. Li. 2020. "Object Detection in UAV Images via Global Density Fused Convolutional Network." *Remote Sensing* 12 (19): 3140. doi:10.3390/rs12193140.

Zhang, S., L. Wen, X. Bian, Z. Lei, and S. Z. Li. 2018. "Single-shot Refinement Neural Network for Object Detection." In *Proceedings of the IEEE conference on computer vision and pattern recognition*. Salt Lake City, Utah, America. June 18–23.

Zhao, Z. Q., P. Zheng, S. T. Xu, and X. Wu. 2019. "Object Detection with Deep Learning: A Review." *IEEE Transactions on Neural Networks and Learning Systems* 30 (11): 3212–3232. doi:10.1109/TNNLS.2018.2876865.

Zhou, L., H. Wei, H. Li, W. Zhao, and Y. Zhang. 2020. "Arbitrary-oriented Object Detection in Remote Sensing Images Based on Polar Coordinates." *IEEE Access* 99: 1.

Zhou, X., C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang. 2017. "East: An Efficient and Accurate Scene Text Detector." In *Proceedings of the IEEE conference on computer vision and pattern recognition*. Honolulu, Hawaii, America. July 21–26.