



## 基于旋转框精细定位的遥感目标检测方法研究

朱煜 方观寿 郑兵兵 韩飞

### Research on Detection Method of Refined Rotated Boxes in Remote Sensing

ZHU Yu, FANG Guan-Shou, ZHENG Bing-Bing, HAN Fei

在线阅读 View online: <https://doi.org/10.16383/j.aas.c200261>

---

## 您可能感兴趣的其他文章

### 基于FlowS-Unet的遥感图像建筑物变化检测

Detection of Building Changes in Remote Sensing Images via FlowS-Unet

自动化学报. 2020, 46(6): 1291–1300 <https://doi.org/10.16383/j.aas.c180122>

### 基于推广流形学习的高分辨遥感影像目标分类

Generalized Manifold Learning for High Resolution Remote Sensing Image Object Classification

自动化学报. 2019, 45(4): 720–729 <https://doi.org/10.16383/j.aas.2017.c170318>

### 改进的YOLO特征提取算法及其在服务机器人隐私情境检测中的应用

An Improved YOLO Feature Extraction Algorithm and Its Application to Privacy Situation Detection of Social Robots

自动化学报. 2018, 44(12): 2238–2249 <https://doi.org/10.16383/j.aas.2018.c170265>

### 基于多尺度先验深度特征的多目标显著性检测方法

Multiple Salient Objects Detection Using Multi-scale Prior and Deep Features

自动化学报. 2019, 45(11): 2058–2070 <https://doi.org/10.16383/j.aas.c170154>

### 基于显著图的弱监督实时目标检测

Weakly Supervised Real-time Object Detection Based on Saliency Map

自动化学报. 2020, 46(2): 242–255 <https://doi.org/10.16383/j.aas.c180789>

### 基于深度学习的桥梁裂缝检测算法研究

Research on Detection Algorithm for Bridge Cracks Based on Deep Learning

自动化学报. 2019, 45(9): 1727–1742 <https://doi.org/10.16383/j.aas.2018.c170052>

# 基于旋转框精细定位的遥感目标检测方法研究

朱煜<sup>1</sup> 方观寿<sup>1</sup> 郑兵兵<sup>1</sup> 韩飞<sup>1</sup>

**摘要** 遥感图像中的目标往往呈现出任意方向排列, 而常见的目标检测算法均采用水平框检测, 并不能满足这类场景的应用需求. 因此提出一种旋转框检测网络 R<sup>2</sup>-FRCNN. 该网络利用粗调与细调两阶段实现旋转框检测, 粗调阶段将水平框转换为旋转框, 细调阶段进一步优化旋转框的定位. 针对遥感图像存在较多小目标的特点, 提出像素重组金字塔结构, 融合深浅层特征, 提升复杂背景下小目标的检测精度. 此外, 为了在金字塔各层中提取更加有效的特征信息, 在粗调阶段设计一种积分与面积插值法相结合的感兴趣区域特征提取方法, 同时在细调阶段设计旋转框区域特征提取方法. 最后在粗调和细调阶段均采用全连接层与卷积层相结合的预测分支, 并且利用 SmoothL<sub>n</sub> 作为网络的回归损失函数, 进一步提升算法性能. 提出的网络在大型遥感数据集 DOTA 上进行评估, 评估指标平均准确率达到 0.7602. 对比实验表明了 R<sup>2</sup>-FRCNN 网络的有效性.

**关键词** 遥感图像, 旋转框检测, 两阶段调整, 像素重组金字塔, 区域特征提取

**引用格式** 朱煜, 方观寿, 郑兵兵, 韩飞. 基于旋转框精细定位的遥感目标检测方法研究. 自动化学报, 2023, 49(2): 415-424

**DOI** 10.16383/j.aas.c200261

## Research on Detection Method of Refined Rotated Boxes in Remote Sensing

ZHU Yu<sup>1</sup> FANG Guan-Shou<sup>1</sup> ZHENG Bing-Bing<sup>1</sup> HAN Fei<sup>1</sup>

**Abstract** The objects in remote sensing images are often shown in any direction. The common algorithms of object detection adopt horizontal detection, which cannot fulfill the application requirements in remote sensing. Therefore, this paper proposes an object detector of rotated boxes named R<sup>2</sup>-FRCNN. The network adopts two stages of rough and refined adjustments to realize the detection of rotated boxes. The rough adjustment stage is used to transform the horizontal boxes into rotated boxes, and the refined adjustment stage is used to further optimize the position of the rotated boxes. In view of the fact that there are many small objects in remote sensing images, this paper proposes a pixel-recombination pyramid structure to improve the detection accuracy of small objects in a complex background by integrating deep and shallow features. In addition, in order to extract more effective feature information from each layer of the pyramid, this paper designs a region pooling method combining integration and area interpolation in the rough adjustment stage, and a region pooling method of rotated boxes in the refined adjustment stage. Finally, this paper adopts the prediction branch combining the fully connected layers and the convolutional layers, and takes the SmoothL<sub>n</sub> as the regression loss function of the network to further improve the performance of the algorithm. The network proposed in this paper is evaluated on a large remote sensing dataset DOTA, and the evaluation mean average precision reaches 0.7602. Comparative experiments show the effectiveness of R<sup>2</sup>-FRCNN modules.

**Key words** Remote sensing images, rotated boxes detection, two stages adjustment, pixel-recombination pyramid, region feature extraction

**Citation** Zhu Yu, Fang Guan-Shou, Zheng Bing-Bing, Han Fei. Research on detection method of refined rotated boxes in remote sensing. *Acta Automatica Sinica*, 2023, 49(2): 415-424

近年来, 随着遥感技术的发展, 高质量的遥感图像日益增多, 这为遥感领域的应用奠定了基础.

遥感图像广泛应用于灾害监测、资源调查、土地利用评价、农业产值测算、城市建设规划等领域<sup>[1]</sup>, 对于社会 and 经济发展具有重要的意义. 而目标检测作为遥感图像处理的应用之一, 获得图中特定目标类别和位置. 通常关注飞机、机场、船舶、桥梁和汽车等目标, 因此对于民用和军用领域有着十分重要的用途<sup>[2]</sup>. 在民用领域中, 船舶的定位有利于海上救援行动, 车辆的定位有利于车辆计数和分析道路的拥堵情况等. 在军事领域中, 这些类别信息的检测获

收稿日期 2020-04-29 录用日期 2020-09-07

Manuscript received April 29, 2020; accepted September 7, 2020

上海市科学技术委员会 (17DZ1100808) 资助

Supported by Shanghai Science and Technology Committee (17DZ1100808)

本文责任编辑 桑农

Recommended by Associate Editor SANG Nong

1. 华东理工大学信息科学与工程学院 上海 200237

1. School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237

取,有利于快速且精准地锁定攻击目标位置、分析战争形势以及制定军事行动等.因此对于遥感图像中的目标进行精准检测至关重要.

目标检测是计算机视觉领域中一个重要且具有挑战性的研究热点.随着深度学习的快速发展,目标检测器的性能取得了显著进步,已经广泛应用于各个行业.目前常用的目标检测器大致可以分为两级检测器和单级检测器两类<sup>[3]</sup>.两级检测器是基于区域卷积神经网络 (Regions with convolutional neural network, R-CNN) 框架,检测过程分为两个阶段.第1阶段从图像中生成一系列候选框区域,第2阶段从候选框区域中提取特征,然后使用分类器和回归器进行预测. Faster R-CNN<sup>[4]</sup> 作为两级检测器的经典方法,提出候选区域生成网络 (Region proposal networks, RPN) 用于候选框的产生,从而快速、准确地实现端到端检测.之后区域全卷积网络 (Region-based fully convolutional network, R-FCN)<sup>[5]</sup>、Cascade R-CNN<sup>[6]</sup> 等两级检测器的出现进一步提高目标检测的精度.单级检测器将检测问题简化为回归问题,仅仅由一系列卷积层进行分类回归,而不需要产生候选框及特征提取阶段.因此这类方法通常检测速度较快.例如,Redmon 等<sup>[7]</sup> 提出 YOLO 检测器,将图像划分为一系列网格区域,每个网格区域直接回归得到边界框. Liu 等<sup>[8]</sup> 提出 SSD 检测器,在多个不同尺度大小的特征图上直接分类回归. Lin 等<sup>[9]</sup> 提出 Focal Loss 分类损失函数,解决单级检测器的类别不平衡问题,进一步提高检测精度.这些先进的目标检测技术往往用于水平边界框的生成,然而在遥感图像中,大多数检测目标呈现出任意方向排列,对于纵横比大或者密集排列的目标,仅仅采用水平框检测将包含过多的冗余信息,影响检测效果.因此旋转方向成为不可忽视的因素.

早期应用于遥感领域的旋转框检测算法主要来源于文本检测,例如 R<sup>2</sup>CNN<sup>[10]</sup> 和 RPN<sup>[11]</sup> 等.然而由于遥感图像背景复杂且空间分辨率变化较大,相比于二分类的文本检测具有更大困难,因此这些优秀的文本检测算法直接应用于遥感领域中并不能取得较好的检测效果.近年来,随着目标检测算法的发展以及针对遥感图像的深入研究,涌现出许多性能良好的旋转框检测算法.例如 Ding 等<sup>[12]</sup> 提出旋转感兴趣区域学习器 (Region of interest transformer, RoI),将水平框转换为旋转框,并在学习器中执行边界框的回归; Zhang 等<sup>[13]</sup> 提出通过捕获全局场景和局部特征的相关性增强特征; Azimi 等<sup>[14]</sup> 提出基于多尺度卷积核的图像级联方法; Yang 等<sup>[15]</sup> 提出像素注意力机制抑制图像噪声,突出目标的特

征,并且在 SmoothL<sub>1</sub> 损失<sup>[4]</sup> 中引入 IoU 常数因子解决旋转框的边界问题,使旋转框预测更加精确. Yang 等<sup>[16]</sup> 设计精细调整模块,采用特征调整模块,通过插值操作实现特征对齐. Xu 等<sup>[17]</sup> 提出回归 4 种长度比来表示对应边的相对偏移距离,并且引入了一个真实框与其水平边界框面积比作为倾斜因子,用于对每个目标水平或旋转检测的选择. Wei 等<sup>[18]</sup> 提出利用预测内部中线实现旋转目标检测的方法. Li 等<sup>[19]</sup> 提出利用预测的掩模获取旋转框的方法. Wang 等<sup>[20]</sup> 提出了一种基于初始横向连接的特征金字塔网络 (Feature pyramid networks, FPN) 增强算法,同时利用语义注意力机制网络提供语义特征,从复杂的背景中提取目标.

因此,目前在遥感图像中用于旋转框检测的方法大致可以分为两种.其中一种算法整体结构仍然为水平框检测,仅仅在回归预测分支中增加一些变量的获取,例如角度因子等.这种算法使得在网络预测的像素中包含较多背景信息,容易出现图 1 所示的角度偏移以及漏检较多等问题.另一种算法预设含有角度的锚点框,然后采用旋转候选框内的像素进行预测.由于目标的旋转角度较多,因此这种算法需要预设大量的锚点框以保证召回率,这样会极大地增加计算量.

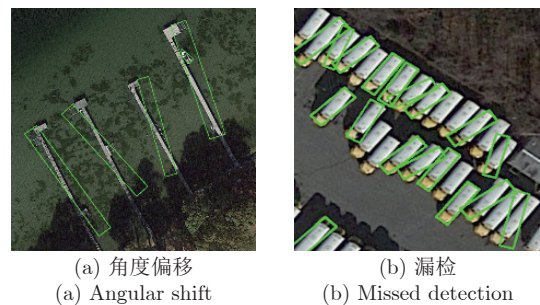


图 1 遥感图像目标检测问题可视化  
Fig. 1 Visualization of remote sensing images object detection problem

针对上述不足,本文结合这两种处理方法的优点,以 Faster R-CNN<sup>[21]</sup> 为基础,提出一种用于旋转框检测的网络 R<sup>2</sup>-FRCNN (Refined rotated faster R-CNN). 该网络依次采用上述两种旋转框处理方法,将前一种方法得到旋转框的过程视为粗调,这个阶段产生的旋转框作为后一种方法的预设框,然后对于旋转框再次进行调整,这个过程称为细调.两阶段调整使得网络输出更加精确的预测框.此外,针对遥感图像存在较多小目标的特点,本文提出像素重组特征金字塔结构 (Pixel-recombination feature pyramid network, PFPN),相比于传统的金字

塔网络, 本文的金字塔结构使得特征局部信息与全局信息相结合, 从而突出复杂背景下小目标的特征响应. 同时为了更好地提取表征目标信息的特征, 用于后续预测阶段, 本文在粗调阶段设计积分感兴趣区域池化方法 (Integrate region of interest pool, IRoIPool), 以及在精调阶段设计旋转感兴趣区域池化方法 (Rotated region of interest pool, RRoIPool), 提升复杂背景下小目标的检测精度. 最后, 本文在粗调和细调阶段均采用全连接层与卷积层结合的预测分支以及 Smooth $L_n$  回归损失函数, 进一步提升算法性能.

本文结构安排如下: 第 1 节详细阐述本文提出的旋转框检测网络 R<sup>2</sup>-FRCNN; 第 2 节通过与官方基准方法和现有方法的实验结果进行对比, 以及本文方法各模块的分离实验, 评估本文方法的性能; 第 3 节总结.

## 1 旋转框目标检测方法

本节对提出的网络 R<sup>2</sup>-FRCNN 结构以及各模块进行阐述. 首先介绍 R<sup>2</sup>-FRCNN 网络的整体结构, 然后详细介绍各个模块 (像素重组金字塔结构、感兴趣区域特征提取和网络预测分支结构), 最后介绍本文使用的损失函数.

### 1.1 网络结构设计

图 2 展示了 R<sup>2</sup>-FRCNN 网络的整体结构, 可以分为基础网络、像素重组金字塔、候选区域生成网络 RPN、粗略调整阶段和精细调整阶段 5 个部分.

本文采用 ResNet<sup>[22]</sup> 作为算法的基础网络, 将  $C_3$ 、 $C_4$ 、 $C_5$  和  $C_6$  特征层用于构建特征金字塔结构, 增强网络对于小目标的检测能力. 由金字塔产生的  $P_3$ 、 $P_4$ 、 $P_5$ 、 $P_6$  和  $P_7$  5 个特征层上, 每个像素点预

设 3 个锚点框, 锚点框的长宽比为 {1:1, 1:2, 2:1}, 尺寸大小为 8, 经由 RPN<sup>[4]</sup> 调整锚点框的位置生成一系列候选框. 然后选择置信度较高的 2000 个候选框用于粗略调整阶段, 该模块的回归过程将水平框调整为旋转框. 最后这些候选框进入精细调整阶段, 再次调整旋转框的位置, 得到更好的检测效果. 经过两阶段调整后的框, 选择后一阶段中最大分类数值作为置信度, 同时采用旋转非极大抑制算法处理, 选取邻域内置信度较高的框, 并且抑制低置信度的框, 这些高置信度的候选框即为网络输出预测框.

### 1.2 像素重组金字塔结构

特征金字塔结构<sup>[23]</sup> 被广泛应用于许多先进的目标检测算法中, 这个结构的设计在于浅层的定位信息准确, 深层的语义信息丰富, 通过融合深层次特征图, 提升对于小目标的检测性能. 如表 1 所示, RoI-Transformer (RT)<sup>[12]</sup>、CADNet<sup>[13]</sup>、SCRDet<sup>[15]</sup>、R<sup>3</sup>Det<sup>[16]</sup> 和 GV R-CNN (GV)<sup>[17]</sup> 均采用了深层次融合特征, 表现出优异的检测性能, 而 R<sup>2</sup>CNN<sup>[10]</sup> 未使用特征融合, 取得的检测结果远低于其他方法. 图 3 为本文设计的像素重组金字塔结构. 该结构分为 2 个阶段: 第 1 阶段为  $C_i \rightarrow M_i$ , 采用尺度转化的方式, 利用局部特征信息的同时, 融合上下层构建金字塔结构; 第 2 阶段为  $M_i \rightarrow P_i$ , 采用非局部注意力<sup>[24]</sup> 模块, 利用全局信息, 突出目标区域的特征.

在第 1 阶段中, 特征上采样对于金字塔结构是一个关键的操作. 最常用的特征上采样方式为插值和转置卷积<sup>[25]</sup>. 插值法仅考虑相邻像素, 无法获取密集预测任务所需的丰富语义信息. 转置卷积作为卷积的逆运算, 将其作为上采样方式存在 2 点不足<sup>[26]</sup>: 1) 对于整个特征图都采用同样的卷积核, 而不考虑特征图中的目标信息, 限制了上采样过程对于局部

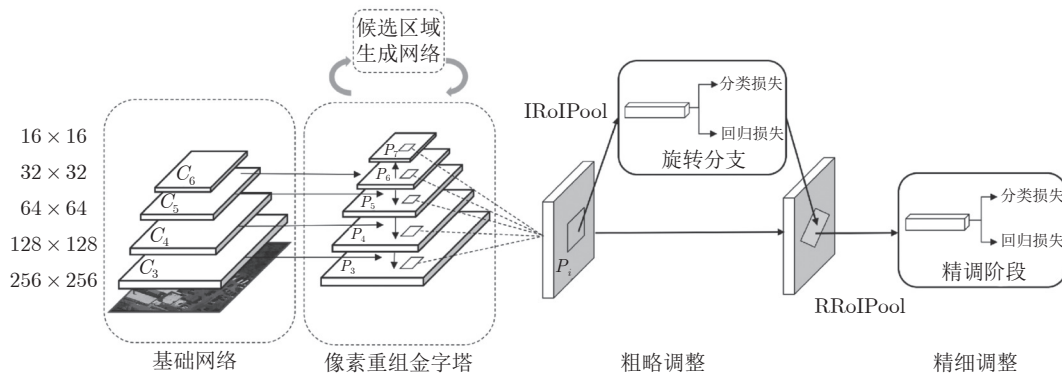


图 2 R<sup>2</sup>-FRCNN 网络结构图

Fig. 2 The structure of R<sup>2</sup>-FRCNN

表 1 不同方法在 DOTA 数据集的检测精度对比 (%)  
Table 1 Comparison of detection accuracy of different methods in DOTA (%)

类别	R <sup>2</sup> CNN <sup>[10]</sup>	RT <sup>[12]</sup>	CADNet <sup>[13]</sup>	SCRDet <sup>[15]</sup>	R <sup>3</sup> Det <sup>[16]</sup>	GV <sup>[17]</sup>	本文方法
飞机	80.94	88.64	87.80	<b>89.98</b>	89.24	89.64	89.10
棒球场	65.67	78.52	82.40	80.65	80.81	<b>85.00</b>	81.22
桥梁	35.34	43.44	49.40	52.09	51.11	52.26	<b>54.47</b>
田径场	67.44	75.92	73.50	68.36	65.62	<b>77.34</b>	72.97
小型车辆	59.92	68.81	71.10	68.36	70.67	73.01	<b>79.99</b>
大型车辆	50.91	73.68	64.50	60.32	76.03	73.14	<b>82.28</b>
船舶	55.81	83.59	76.60	72.41	78.32	86.82	<b>87.64</b>
网球场	90.67	90.74	<b>90.90</b>	90.85	90.83	90.74	90.54
篮球场	66.92	77.27	79.20	<b>87.94</b>	84.89	79.02	87.31
储油罐	72.39	81.46	73.30	<b>86.86</b>	84.42	86.81	86.33
足球场	55.06	58.39	48.40	65.02	<b>65.10</b>	59.55	54.20
环形车道	52.23	53.54	60.90	66.68	57.18	<b>70.91</b>	68.18
港口	55.14	62.83	62.00	66.25	68.10	72.94	<b>76.12</b>
游泳池	53.35	58.93	67.00	68.24	68.98	<b>70.86</b>	70.83
直升机	48.22	47.67	62.20	<b>65.21</b>	60.88	57.32	59.19
平均准确率	60.67	69.56	69.90	72.61	72.81	75.02	<b>76.02</b>

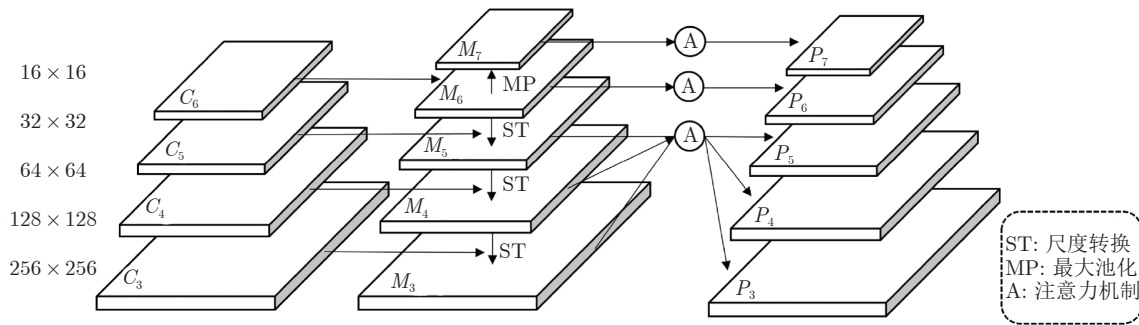


图 3 像素重组金字塔结构

Fig.3 The structure of pixel-recombination pyramid

变化的响应; 2) 若采用较大的卷积核将会增加大量参数. 本文引入尺度转换作为特征上采样方法. 深浅层特征融合的操作过程如图 4 所示. 该方法首先利用“通道转化”方法<sup>[27]</sup>压缩通道数 (本文压缩系数  $r = 0.5$ ), 增大特征图尺寸, 即:

$$I_{H,W,C} = I_{\lfloor H/r \rfloor, \lfloor W/r \rfloor, C \cdot r^2 + r \cdot \text{mod}(W,r) + \text{mod}(H,r)} \quad (1)$$

然后, 采用  $1 \times 1$  的卷积层用于调整通道数, 再由 Softmax 函数<sup>[28]</sup>作用于每一通道的特征层. 最后采用式 (2) 进行加权求和, 使得特征融合过程更好地利用局部信息.

$$\begin{cases} y_{m,n,c} = \sum_{i=-2}^2 \sum_{j=-2}^2 x_{m+i,n+j,c} \cdot w_{m,n,k} \\ k = (i+2) \times 5 + j + 2 \end{cases} \quad (2)$$

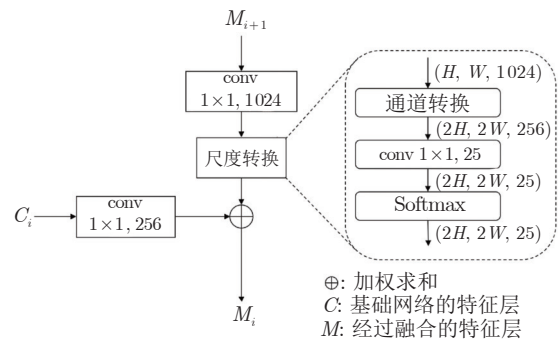


图 4 特征融合结构

Fig.4 The structure of feature fusion

式中,  $m$ 、 $n$  分别表示像素的横、纵位置,  $c$  表示  $C$  特征层当前通道,  $k$  表示  $M$  特征层当前通道.

第 2 阶段采用非局部注意力模块, 利用特征图中目标与全局特征的关系, 突出目标区域的响应.

根据非局部注意力模块的定义, 假设  $C$  为通道数,  $s$  为尺度大小,  $G$  为特征图尺度的乘积即  $s \times s$ ,  $x$  为输入特征图,  $q(x)$ 、 $k(x)$  和  $v(x)$  定义为采用不同线性转换的结果:

$$q(x^s) = W_q^s \mathbf{T} x^s \quad (3)$$

$$k(x^s) = W_k^s \mathbf{T} x^s \quad (4)$$

$$v(x^s) = W_v^s \mathbf{T} x^s \quad (5)$$

式中, 系数矩阵  $W_q^s$ ,  $W_k^s \in \mathbf{R}^{C \times C/8}$ ,  $W_v^s \in \mathbf{R}^{C \times C}$ .

$q(x^s)$  与  $k(x^s)$  矩阵相乘, 得二维矩阵  $o^s \in \mathbf{R}^{G \times G}$ ; 再运用 Softmax 将矩阵的每一行转换为概率值, 最后与  $v(x^s)$  矩阵相乘后再与输入相加, 得输出量  $x^{s'}$ :

$$x^{s'} = x^s + (o^s v^s)^{\mathbf{T}} \quad (6)$$

在本文的特征金字塔结构中, 第 1 阶段输出的  $M_3$  和  $M_4$  由于尺度较大, 直接用于非局部注意力模块计算量较大. 因此为了保留这两层的语义信息, 同时再次融合不同层的特征, 该结构将  $M_3$  和  $M_4$  池化为  $M_5$  的尺寸大小, 然后计算这 3 层的均值输入非局部注意力模块, 再由插值操作输出对应相等尺寸的特征图.  $M_6$  和  $M_7$  的特征图直接应用非局部注意力模块得到  $P_6$  和  $P_7$  层.

### 1.3 感兴趣区域特征提取模块

感兴趣区域特征提取模块主要用于固定输出尺寸大小, 提取表征框内区域的特征, 便于后续的网络预测. 本文的 RoI 特征提取模块主要分为粗调阶段的水平框和细调阶段的旋转框 RoI 特征提取两部分.

自然场景图像中的目标通常是固定方向呈现, 因此两阶段式目标检测算法采用水平框的 RoI 特征提取. 目前, 应用较为广泛的 RoI 特征提取是 RoIPooling<sup>[4]</sup> 和 RoI Align<sup>[29]</sup>. 图 5(a) 为 RoI 池化原理图, 选择量化后块中最大像素值作为池化后的结果. 然而量化的结果会导致提取的小目标像素存在偏差, 影响检测效果. 图 5(b) 为 RoI 对齐原理图, 取消量化操作, 采用双线性插值在块中计算出  $N$  个浮点坐标的像素值, 均值作为块的结果. 然而这个操作存在两点不足: 采样点数量需要预先设置, 不同大小候选框设置了相同数量的采样点.

因此, 本文采用精确 RoI (Precise RoI, Pr-RoI) 池化方法<sup>[30]</sup> 的特征提取操作, 如图 6 所示, 由插值操作将块内特征视为一个连续的过程, 采用积分方

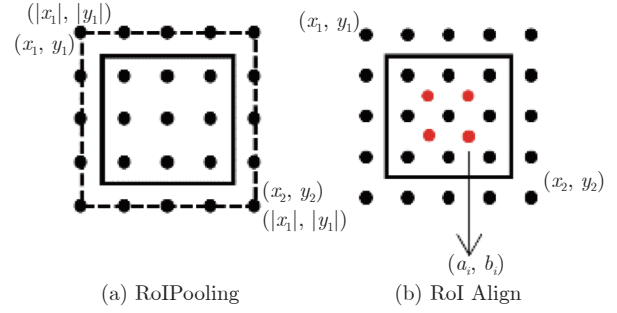


图 5 常用 RoI 特征提取示意图

Fig. 5 The schematic diagram of common RoI feature extraction

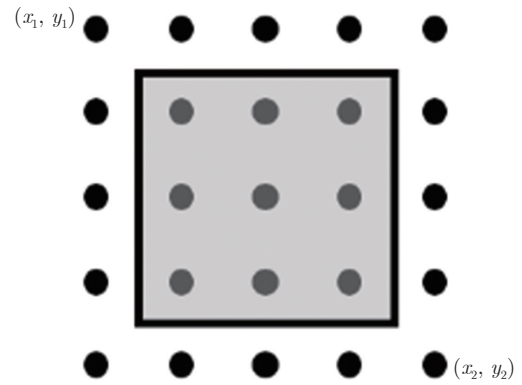


图 6 IRoIPool 特征提取示意图

Fig. 6 The diagram of IRoIPool feature extraction

法获得整个块的像素和, 其均值作为块的结果, 即:

$$\text{IRoIPool}(bin, \mathcal{F}) = \frac{\int_{y_1}^{y_2} \int_{x_1}^{x_2} f(x, y) dx dy}{(x_2 - x_1) \times (y_2 - y_1)} \quad (7)$$

式中,  $f(x, y)$  为采用面积插值法<sup>[15]</sup> 所得的像素值.

旋转框 RoI 特征提取直接采用积分操作较为复杂, 因此本文将积分操作视为块内一定数量的像素之和, 从而得到块的均值, 即:

$$\text{RRoIPool}(bin, \mathcal{F}) = \frac{\sum_{y=y_1}^{y_2} \sum_{x=x_1}^{x_2} f(x, y)}{N_x \times N_y} \quad (8)$$

$$N_x = \left\lfloor \frac{x_2 - x_1}{l_x} \right\rfloor + 1, N_y = \left\lfloor \frac{y_2 - y_1}{l_y} \right\rfloor + 1 \quad (9)$$

式中,  $(x_1, y_1)$  和  $(x_2, y_2)$  分别为旋转框在水平位置处的左上角和右下角点,  $l_x$  和  $l_y$  分别为水平方向和垂直方向的采样距离, 如图 7 所示.

根据候选框的大小决定采样点的数量. 然而采样距离太小会导致计算量大幅增加, 因此为平衡检测效率与精度, 本文将采样距离  $l_x$  和  $l_y$  设置为 0.4.

旋转框在水平位置处采样点的坐标为  $(x_h, y_h)$ ,

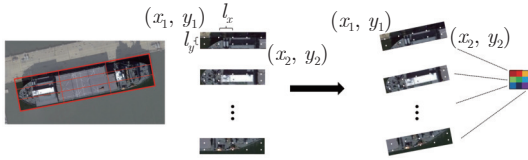


图 7 旋转 RoI 特征提取示意图

Fig.7 The diagram of rotated RoI feature extraction

旋转框  $w$  所对应的边与横轴正方向的夹角为  $\theta$ , 旋转框的中心点为  $(c_x, c_y)$ , 由式 (10) 转化为旋转框中的坐标  $(x, y)$ , 再由面积插值法得到该位置的像素值.

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta & (1-\cos\theta)\cdot c_x + \sin\theta\cdot c_y \\ \sin\theta & \cos\theta & -\sin\theta\cdot c_x + (1-\cos\theta)\cdot c_y \end{bmatrix} \begin{bmatrix} x_h \\ y_h \\ 1 \end{bmatrix} \quad (10)$$

本文方法与 R<sup>3</sup>Det 类似, 都使用了精细调整旋转框的定位. 然而 R<sup>3</sup>Det 每一次调整的预测分支直接采用卷积层操作, 但是卷积操作为水平滑动, 用于旋转框回归将会包含一些背景像素干扰预测结果, 而本文方法采用旋转框感兴趣区域提取框内的特征信息用于预测, 更加有利于检测性能的提升.

#### 1.4 预测分支结构

目标检测算法分为定位和分类两个任务. 一般而言, 两级检测器的预测分支采用全连接层, 而单级检测器的预测分支采用卷积层. Wu 等<sup>[31]</sup> 发现这两个任务适合于不同的预测分支结构, 全连接层更适合用于分类任务, 卷积层更适合用于回归任务. 因此, 本文采用图 8 所示的预测分支结构.

在本文采用的预测分支中, 分类结构保持不变, 仍然采用全连接层. 而回归分支采用一系列 ResNet 网络中的 ResBlock 结构 (本文使用 2 个).

#### 1.5 网络训练损失函数

本文提出网络的损失函数包含 RPN 阶段  $L_{RPN}$ 、粗略调整阶段  $L_{ro}$  和精细调整阶段  $L_{re}$ , 即:

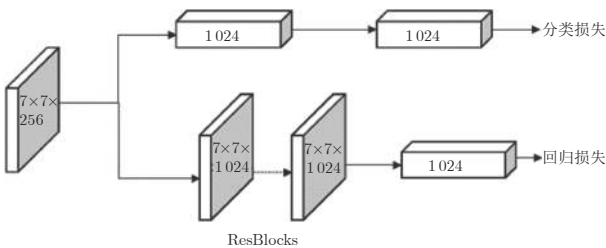


图 8 预测分支结构图

Fig.8 The diagram of prediction branch

$$L = L_{RPN} + L_{ro} + L_{re} \quad (11)$$

每一阶段的损失函数都包含分类损失和回归损失. 分类损失采用交叉熵损失函数<sup>[4]</sup>. 回归损失采用 Smooth $L_n$  损失函数<sup>[32]</sup>, 如式 (12) 所示, 相比于 Smooth $L_1$  损失函数<sup>[4]</sup>, 该损失函数的一阶导数是连续存在的, 具有良好的光滑性.

$$SL_n(x) = (|x| + 1) \ln(|x| + 1) - |x| \quad (12)$$

$$\frac{\partial SL_n(x)}{\partial x} = \text{sign}(x) \cdot \ln(\text{sign}(x) \cdot x + 1) \quad (13)$$

此外, 式 (11) 中 RPN 阶段为水平框的回归, 因此使用  $x, y, w, h$  4 个值代表水平框. 粗调阶段和细调阶段为旋转框的回归, 使用  $x, y, w, h, \theta$  5 个值代表旋转框, 因此旋转框的回归转换值定义为:

$$\begin{bmatrix} t_x \\ t_y \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x_t - x_a \\ y_t - y_a \end{bmatrix} \begin{bmatrix} \frac{1}{w_a} & 0 \\ 0 & \frac{1}{h_a} \end{bmatrix} \quad (14)$$

$$t_w = \log_2\left(\frac{w_t}{w_a}\right), \quad t_h = \log_2\left(\frac{h_t}{h_a}\right) \quad (15)$$

$$t_\theta = (\theta_t - \theta_a) \bmod 2\pi \quad (16)$$

式中,  $x, y, w, h, \theta$  分别为旋转框中心点的横、纵坐标, 框的宽度、高度和旋转角度.  $x_t, x_a$  分别表示真实框和候选框的值.

## 2 实验结果与分析

本文实验设备使用英特尔 E5-2683 CPU, 英伟达 GTX 1080Ti 显卡, 64 GB 内存的服务器, 实验环境为 Ubuntu 16.04.4 操作系统、Cuda9.0、Cudnn7.4.2、Pytorch1.1.0、Python3.7.

本文实验中采用 3 个 GPU 进行训练, 批处理大小为 3 (GPU 显存限制), 输入图像统一为 1024 × 1024 分辨率. 训练的迭代次数为 15 轮, 同时使用衰减系数为 0.0001、动量为 0.9 的随机梯度下降作为优化器, 初始的学习率设置为 0.01, 分别在第 8、第 11 轮和第 14 轮将学习率降低 10 倍. 图 9 是在 DOTA 数据集上训练过程的损失下降曲线图 (一轮训练有 4500 次迭代), 在第 8 轮 (36000 次迭代) 出现明显的损失下降.

### 2.1 实验数据集

本文使用 DOTA<sup>[21]</sup> 用于算法的评估. DOTA 是由旋转框标注的大型公开数据集, 主要用于遥感图像目标检测任务. 该数据集包含由各个不同传感

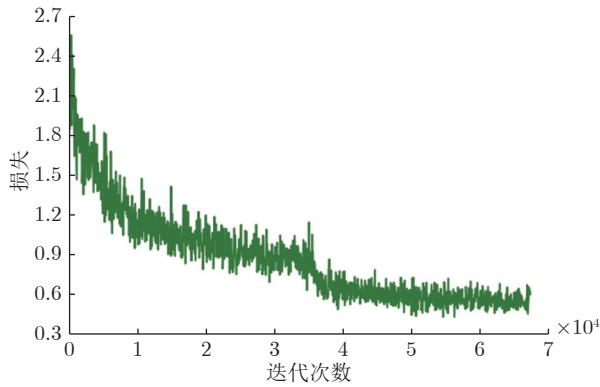


图9 在DOTA上训练过程损失曲线图

Fig.9 Train loss on DOTA

器和平台采集的 2806 张图像, 图像的大小范围从  $800 \times 800$  像素到  $4000 \times 4000$  像素, 含有各种尺度、方向和形状. 专家选择 15 种常见类别对这些图像进行标注, 总共标注 188282 个目标对象, 包括飞机、棒球场、桥梁、田径场、小型车辆、大型车辆、船舶、网球场、篮球场、储油罐、足球场、环形车道、港口、游泳池和直升机. 另外该数据集选取一半的图像作为训练集, 1/6 作为验证集, 1/3 作为测试集, 其中测试集的标注不公开. 为降低高分辨率图像由于压缩对于小目标的影响, 本文将所有图像统一裁剪为  $1024 \times 1024$  的子图像, 重叠为 200 像素.

## 2.2 检测结果对比

本文方法采用 ResNet50 与可变形卷积<sup>[33]</sup> 相结合作为基础网络进行本节实验. 为了评估本文方法的性能, 实验数据均采用官方提供的训练集和测试集. 实验结果通过提交到 DOTA 评估服务器上获得, 本文方法的评估结果平均准确率为 0.7602, 超过目前官方提供的基准方法<sup>[21]</sup>.

除了与官方基准方法进行对比, 本节实验还与  $R^2$ CNN<sup>[10]</sup>、RoI-Transformer<sup>[12]</sup>、CADNet<sup>[13]</sup>、SCRDet<sup>[15]</sup>、 $R^3$ Det<sup>[16]</sup> 和 GV R-CNN<sup>[17]</sup> 进行对比分析, 各方法的检测结果如表 1 所示.

由表 1 中的检测结果可以看出, 本文方法的检测结果优于其他方法, 达到 76.02% 的平均准确率. 其中桥梁、小型车辆、大型车辆、船舶和港口这些类别取得最高检测精度. 由图 10 可以看出, 这些类别的目标在遥感数据集中尺寸较小, 并且往往呈现出密集排列, 因此说明本文方法对于在这类场景的检测更具有优势. 此外, 飞机、网球场、篮球场、储水池、游泳池等类别在遥感数据集中尺寸较大, 对于这些目标本文方法仍取得与其他方法中最高检测精度相差不大的结果. 这些检测结果说明本文方法能够有效用于检测遥感图像中的目标.

## 2.3 分离实验

### 1) 各模块对于检测精度的影响

为验证本文方法各模块的有效性, 本节进行了一系列对比实验. 表 2 展示了网络在 DOTA 数据集上不同模块设置的检测结果. 其中“√”表示采用该项设置, ConvFc 表示采用第 1.4 节设计的预测分支结构. 对比实验分析如下:

a) 基准设置. 本节实验将扩展后的 Faster R-CNN OBB<sup>[21]</sup> 用于旋转框检测任务. 其中, 基础网络采用 ResNet50<sup>[22]</sup>, 并且采用特征金字塔<sup>[23]</sup>, RoI 特征提取采用 RoI Align<sup>[29]</sup>, 回归分支采用 Smooth $L_1$  损失函数<sup>[4]</sup>. 为了保证实验的公平性和准确性, 后续实验参数设置都是严格一致.

b) 精细调整. 在实验的精细调整阶段, 初始候选区域特征提取选择 Rotated RoI Align (RRoI Align) 方法, 该方法为 RoI Align<sup>[29]</sup> 在旋转框中的应用. 由表 2 的结果显示, 精细调整阶段的添加, 使得检测效果得到大幅提升, 评估指标平均准确率增加 4.10%. 说明提取旋转候选框内像素进一步调整是有必要的, 这个阶段避免了水平框特征提取包含过多背景像素的问题, 从而提升对较大纵横比目标的检测效果. 然而在实验中发现, 在精细调整结构中多次调整提升效果并不明显, 从一次调整增加为两次调整, 平均准确率为 73.68%, 仅仅增加 0.06%, 因此为了减少参数量, 本文后续实验的精细调整阶段采用一次调整过程.

c) RoI 特征提取. 实验中, 将第 1.3 节提出的 IRoIPool 和 RRoIPool 用于替换初始两阶段调整模块的 RoI Align 和 RRoI Align. 由表 2 的实验结果显示, 相比于初始 RoI 特征提取方法, IRoIPool 方法使得检测精度平均准确率提升 0.37%, RRoIPool 方法使得检测精度平均准确率进一步提升 0.32%, 说明本文设计的 RoI 特征提取更为有效. 本文后续将对这两个特征提取方法的结构做进一步研究.

d) PFPN 结构. 为了更好地验证 PFPN 的作用, 本文对此设计了两组实验. 第 1 组, 金字塔结构的深浅层不进行尺寸转化和非局部注意力模块, 仅仅采用  $1 \times 1$  的卷积将特征层的通道数转化为 256, 网络的其他结构和训练超参数保持一致, 平均准确率仅为 64.55%, 由于 DOTA 数据集中小目标较多, 因此说明 PFPN 金字塔结构对于小目标的检测效果显著. 第 2 组实验的结果见表 2, 相比于 FPN, PFPN 使得平均准确率提升 0.66%, 说明本文提出的 PFPN 结构对于遥感目标的检测更为有效.

e) 网络预测分支. 本节针对预测分支进行两部分的实验, 即回归损失函数和预测分支结构. 由表 2

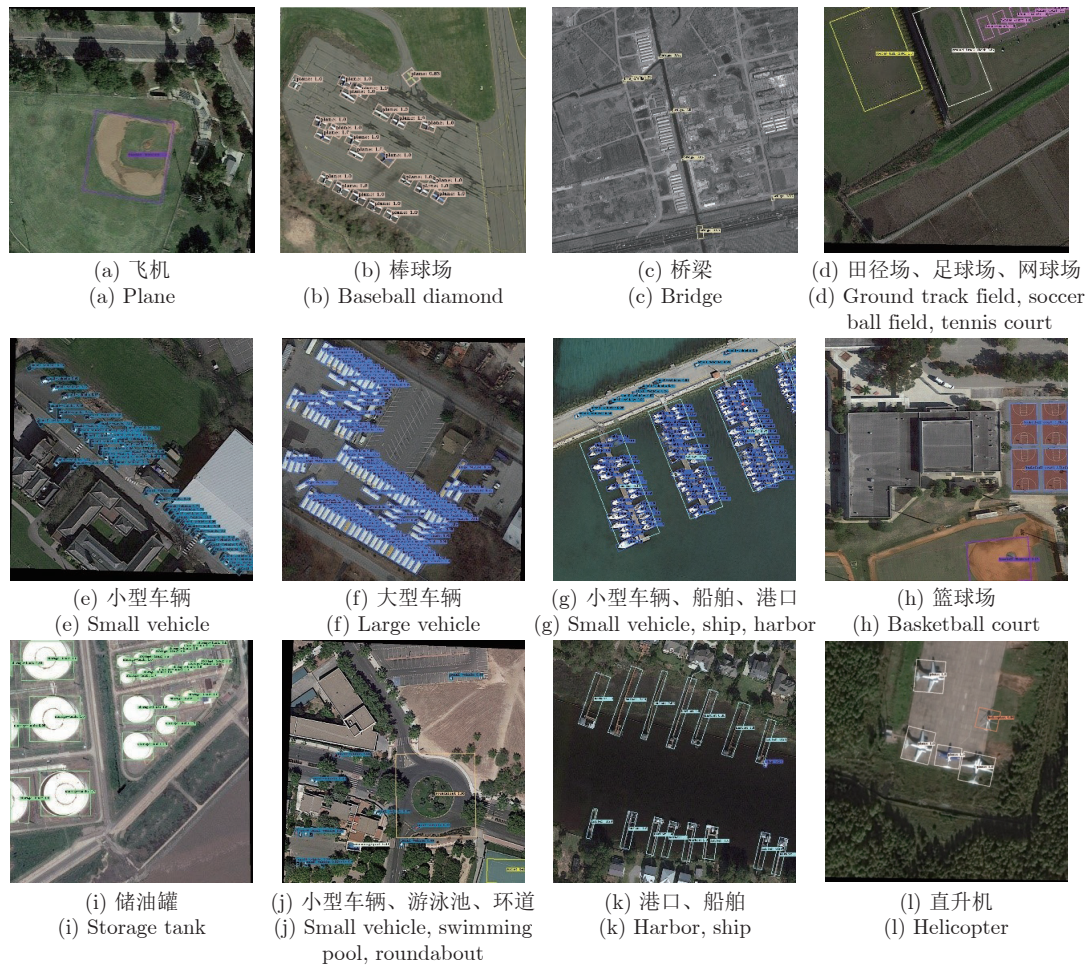


图 10 各类别检测结果展示

Fig.10 Visualization of each category detection

表 2 R<sup>2</sup>-FRCNN 模块分离检测结果Table 2 R<sup>2</sup>-FRCNN module separates detection results

模块	R <sup>2</sup> -FRCNN						
基准设置	✓	✓	✓	✓	✓	✓	✓
精细调整		✓	✓	✓	✓	✓	✓
IRoIPool			✓	✓	✓	✓	✓
RRoIPool				✓	✓	✓	✓
PFPN					✓	✓	✓
SmoothL <sub>n</sub>						✓	✓
ConvFc							✓
平均准确率 (%)	69.52	73.62	73.99	74.31	74.97	75.13	75.96

可以看出, 相比于 SmoothL<sub>1</sub>, 回归损失函数采用 SmoothL<sub>n</sub>, 使得检测精度平均准确率提升 0.16%。此外, 采用第 1.4 节所设计的预测分支结构, 分类过程采用全连接层, 回归过程采用卷积层, 仅增加 2 个 ResBlock 模块, 使得平均准确率提升 0.83%。由此说明回归过程采用 SmoothL<sub>n</sub> 函数和卷积层更加适合旋转框目标检测。

## 2) 感兴趣区域特征提取模块研究

本节研究不同 RoI 特征提取结构对于检测精度的影响, 实验分为水平候选框特征提取方法和旋转候选框特征提取方法两部分。实验结果分别见表 3 和表 4 所示。

表 3 的实验结果显示, 采用 RoIPooling 方式的检测精度相对较低, 其量化操作降低了对于小目标

表 3 不同水平框特征提取方法的实验结果  
Table 3 Experimental results of feature extraction methods of different horizontal boxes

模块	平均准确率 + 精细调整		
	方法	RoIPooling	RoI Align
平均准确率 (%)	71.21	73.62	73.99

表 4 不同旋转框特征提取方法的实验结果  
Table 4 Experimental results of different feature extraction methods of rotated boxes

模块	平均准确率 + 精细调整 + IRoIPool		
	方法	RRoI A-Pooling	RRoI Align
平均准确率 (%)	73.38	73.99	74.31

的检测效果. 而 RoI Align 方式取消量化操作, 采用插值方式使得平均准确率提升 2.41%, 说明提取连续的特征有利于目标检测. 本文方法在面积插值法的基础上引入积分操作, 平均准确率提升 0.37%. 相比于前一种方式选取固定数量的像素点, 本文采用的积分操作类似于选取较多点, 可以提取更多特征, 有利于检测效果的提升.

表 4 为采用不同旋转框特征提取方法的检测结果. 第 1 种方法旋转感兴趣区域平均池化方法 (Rotated region of interest average pooling, RRoI A-Pooling) 选取旋转框内的像素点, 像素均值作为提取的特征. 第 2 种方法采用类似 RoI Align 的方式在旋转框内选择浮点数坐标, 运用双线性插值获得对应的像素值, 平均准确率提升 0.61%. 本文采用方法 RRoIPool 可以根据旋转框大小选择不同数量的像素点表示特征. 相比于第 2 种方式提升 0.32%, 说明本文采用的旋转框特征提取方式更适合于精细调整模块.

### 3 结束语

基于深度学习的目标检测算法在自然场景图像中取得了很大进展. 然而遥感图像存在背景复杂、小目标较多、排列方向任意等难点, 常见的目标检测算法并不满足这类场景的应用需求. 因此本文提出一种粗调与细调两阶段结合的旋转框检测网络 R<sup>2</sup>-FRCNN 用于遥感图像检测任务. 并且设计像素重组金字塔结构, 提高复杂背景下小目标的检测性能. 同时在粗调阶段设计一种水平框特征提取方法 IRoIPool, 细调阶段设计旋转框特征提取方法 RRoIPool. 此外, 本文还采用 SmoothL<sub>n</sub> 回归损失函数, 以及全连接层和卷积层结合的预测分支, 进一步提升检测精度. 实验结果表明本文方法在大型公共数据集 DOTA 上获得了较好的检测效果. 然

而本文方法存在检测速度较慢、GPU 资源消耗较大等缺点, 因此在后续的工作中也将针对网络的轻量化展开进一步研究.

### References

- 1 Ya Y, Pan H, Jing Z L, Ren X G, Qiao L F. Fusion object detection of satellite imagery with arbitrary-oriented region convolutional neural network. *Aerospace Systems*, 2019, 2(2): 163–174
- 2 Wang Yan-Qing, Ma Lei, Tian Yuan. State-of-the-art of ship detection and recognition in optical remotely sensed imagery. *Acta Automatica Sinica*, 2011, 37(9): 1029–1039 (王彦清, 马雷, 田原. 光学遥感图像舰船目标检测与识别综述. 自动化学报, 2011, 37(9): 1029–1039)
- 3 Zhang Hui, Wang Kun-Feng, Wang Fei-Yue. Advances and perspectives on applications of deep learning in visual object detection. *Acta Automatica Sinica*, 2017, 43(8): 1289–1305 (张慧, 王坤峰, 王飞跃. 深度学习在目标视觉检测中的应用进展与展望. 自动化学报, 2017, 43(8): 1289–1305)
- 4 Ren S Q, He K M, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, 39(6): 1137–1149
- 5 Dai J F, Li Y, He K M, Sun J. R-FCN: Object detection via region-based fully convolutional networks. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona, Spain: 2016. 379–387
- 6 Cai Z W, Vasconcelos N. Cascade R-CNN: Delving into high quality object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 6154–6162
- 7 Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016. 779–788
- 8 Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C Y, et al. SSD: Single shot MultiBox detector. In: Proceedings of the 14th European Conference on Computer Vision. Amsterdam, Netherlands: 2016. 21–37
- 9 Lin T Y, Goyal P, Girshick R, He K M, Dollár P. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020, 42(2): 318–327
- 10 Jiang Y Y, Zhu X Y, Wang X B, Yang S L, Li W, Wang H, et al. R<sup>2</sup>CNN: Rotational region CNN for orientation robust scene text detection [Online], available: <https://arxiv.org/abs/1706.09579>, June 29, 2017
- 11 Ma J Q, Shao W Y, Ye H, Wang L, Wang H, Zheng Y B, et al. Arbitrary-oriented scene text detection via rotation proposals. *IEEE Transactions on Multimedia*, 2018, 20(11): 3111–3122
- 12 Ding J, Xue N, Long Y, Xia G S, Lu Q K. Learning RoI transformer for oriented object detection in aerial images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019. 2844–2853
- 13 Zhang G J, Lu S J, Zhang W. CAD-Net: A context-aware detection network for objects in remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 2019, 57(12): 10015–10024
- 14 Azimi S M, Vig E, Bahmanyar R, Körner M, Reinartz P. Towards multi-class object detection in unconstrained remote sensing imagery. In: Proceedings of the 14th Asian Conference on Computer Vision. Perth, Australia: 2019. 150–165
- 15 Yang X, Yang J R, Yan J C, Zhang Y, Zhang T F, Guo Z, et al. SCRDet: Towards more robust detection for small, cluttered and rotated objects. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, South Korea: IEEE, 2019. 8231–8240

- 16 Yang X, Yan J C, Feng Z N, He T. R<sup>3</sup>DET: Refined single-stage detector with feature refinement for rotating object. In: Proceedings of the 35th AAAI Conference on Artificial Intelligence. Virtual Event: 2021. 3163–3171
- 17 Xu Y C, Fu M T, Wang Q M, Wang Y K, Chen K, Xia G S, et al. Gliding vertex on the horizontal bounding box for multi-oriented object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019, **43**(4): 1452–1459
- 18 Wei H R, Zhang Y, Cheng Z H, Li H, Wang H Q, Sun X. Oriented objects as pairs of middle lines [Online], available: <https://arxiv.org/abs/1912.10694>, December 23, 2019
- 19 Li Y Y, Huang Q, Pei X, Jiao L C, Shang R H. RADet: Refine feature pyramid network and multi-layer attention network for arbitrary-oriented object detection of remote sensing images. *Remote Sensing*, 2020, **12**(3): 389–409
- 20 Wang J W, Ding J, Guo H W, Cheng W S, Pan T, Yang W. Mask OBB: A semantic attention-based mask oriented bounding box representation for multi-category object detection in aerial images. *Remote Sensing*, 2019, **11**(24): 2930–2951
- 21 Xia G S, Bai X, Ding J, Zhu Z, Belongie S, Luo J B, et al. DOTA: A large-scale dataset for object detection in aerial images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 3974–3983
- 22 He K M, Zhang X Y, Ren S Q, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016. 770–778
- 23 Lin T Y, Dollár P, Girshick R, He K M, Hariharan B, Belongie S. Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE, 2017. 936–944
- 24 Yi J R, Wu P X, Metaxas D N. ASSD: Attentive single shot multi-box detector. *Computer Vision and Image Understanding*, 2019, **189**: Article No. 102827
- 25 Zeiler M D, Krishnan D, Taylor G W, Fergus R. Deconvolutional networks. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco, USA: IEEE, 2010. 2528–2535
- 26 Wang J Q, Chen K, Xu R, Liu Z W, Loy C C, Lin D. CARAFE: Content-aware reassembly of features. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, South Korea: IEEE, 2019. 3007–3016
- 27 Zhou P, Ni B B, Geng C, Hu J G, Xu Y. Scale-transferrable object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 528–537
- 28 Bridle J S. Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. *Neurocomputing: Algorithms, Architectures and Applications*, 1990: 227–236
- 29 He K M, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017. 2980–2988
- 30 Jiang B R, Luo R X, Mao J Y, Xiao T T, Jiang Y N. Acquisition of localization confidence for accurate object detection. In: Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: 2018. 816–832
- 31 Wu Y, Chen Y P, Yuan L, Liu Z C, Wang L J, Li H Z, et al. Rethinking classification and localization for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE, 2020. 10183–10192
- 32 Liu Y L, Jin L W. Deep matching prior network: Toward tighter multi-oriented text detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE, 2017. 3454–3461

- 33 Dai J F, Qi H Z, Xiong Y W, Li Y, Zhang G D, Hu H, et al. Deformable convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017. 764–773



**朱煜** 华东理工大学信息科学与工程学院教授。1999年获得南京大学博士学位。主要研究方向为图像处理, 计算机视觉, 多媒体通信和深度学习。本文通信作者。

E-mail: zhuyu@ecust.edu.cn

(**ZHU Yu** Professor at the School of Information Science and Engineering, East China University of Science and Technology. She received her Ph.D. degree from Nanjing University in 1999. Her research interest covers image processing, computer vision, multi-media communication, and deep learning. Corresponding author of this paper.)



**方观寿** 华东理工大学信息科学与工程学院硕士研究生。主要研究方向为目标检测, 深度学习。

E-mail: y30180616@mail.ecust.edu.cn

(**FANG Guan-Shou** Master student at the School of Information Science and Engineering, East China University of Science and Technology. His research interest covers object detection and deep learning.)



**郑兵兵** 华东理工大学信息科学与工程学院博士研究生。主要研究方向为医学图像处理, 深度学习和计算机视觉。E-mail: bostonkg@outlook.com

(**ZHENG Bing-Bing** Ph.D. candidate at the School of Information Science and Engineering, East China

University of Science and Technology. His research interest covers medical image processing, deep learning, and computer vision.)



**韩飞** 华东理工大学信息科学与工程学院硕士研究生。主要研究方向为目标检测, 计算机视觉和深度学习。

E-mail: fei-han\_huali@163.com

(**HAN Fei** Master student at the School of Information Science and Engineering, East China University of Science and Technology. His research interest covers object detection, computer vision, and deep learning.)