Software Engineering and Applications 软件工程与应用, 2025, 14(2), 484-498
Published Online April 2025 in Hans. https://www.hanspub.org/journal/sea
https://doi.org/10.12677/sea.2025.142043

Hans汉斯

基于信息扩大集合和自适应特征融合的遥感目标检测

王文聪，张孙杰

上海理工大学光电信息与计算机工程学院，上海

收稿日期：2025年3月21日；录用日期：2025年4月19日；发布日期：2025年4月29日

摘　要

在现有方法的基础上，本文提出了一种新颖的信息扩大集合和自适应特征融合的检测算法。文中在主干网络部分引入ConvNeXt模块来加强对被遮蔽目标的检测能力，提出了信息扩大集合模块来充分地利用图像中的上下文信息，优化对长宽比较大目标的检测效果。使用协调注意力模块来防止目标位置信息的丢失，通过挤压与激励模块对通道重新进行权重分配，挑选出重要性较高的通道进行计算。文中还使用自适应空间特征融合模块，对不同层级的特征图进行融合来保证金字塔的效果。在DOTA-v1.5遥感数据集上，相较于原始网络，文中方法mAP@0.5性能提升了2.5个百分点。在另外的2个数据集上，本文提出的算法也取得了更好的检测性能。

关键词

遥感目标检测，多尺度特征，上下文信息提取，注意力机制，YOLO，小目标检测，特征融合

Remote Sensing Target Detection Based on Information Expansion Collection and Adaptive Feature Fusion

Wencong Wang, Sunjie Zhang

School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai

Received: Mar. 21st, 2025; accepted: Apr. 19th, 2025; published: Apr. 29th, 2025

Abstract

On the basis of existing methods, this paper proposes a novel detection algorithm based on

文章引用: 王文聪, 张孙杰. 基于信息扩大集合和自适应特征融合的遥感目标检测[J]. 软件工程与应用, 2025, 14(2): 484-498. DOI: 10.12677/sea.2025.142043

Information Expansion Collection and Adaptive Feature Fusion. The ConvNeXt module is introduced in the backbone network part to enhance the detection ability of occluded targets; the Information Expansion Collection module is proposed to fully utilize the contextual information in the image to optimize the detection effect of targets with large aspect ratio; and the Coordinated Attention module is used to prevent the loss of target position information. The Squeezing and Excitation module is used to re-assign weights to the channels and select the channels with higher importance for computation; the Adaptive Spatial Feature Fusion module is used to fuse the feature maps of different layers to ensure the effect of pyramid. Finally, on the DOTA-v1.5 remote sensing dataset, compared with the original network, the mAP@0.5 performance improvement of 2.5% was obtained. The algorithm proposed in this paper also achieved better detection performance compared to the current state-of-the-art detection algorithms on the other 2 datasets.

Keywords

Remote Sensing Target Detection, Multi-Scale Feature, Contextual Information Extraction, Attention Mechanisms, YOLO, Small Target Detection, Feature Fusion

1. 引言

目标检测[1]作为图像处理任务当中最重要的任务之一，在现实生活中已经开始广泛的运用。近年来，随着航空航天技术和传感器系统的发展，遥感图像的清晰度也逐步提高。遥感图像当中的对象一般为飞机、车辆、舰船、操场、建筑等多种目标，可以用于国土环境监测[2]、灾害救援[3]、农作物监测[4]、军事领域[5]、故障检测[6]以及其他领域。随着机器视觉研究的不断深入，深度学习方法为提升自然图像和遥感图像的检测效果提供了新的方法和思路。

在自然图像领域，出现了一些经典的深度学习方法，主要分为一阶段方法和两阶段方法。一阶段法包括 SSD (Single Shot Multibox Detector) [7]、RetinaNet [8]和 YOLO (You Only Look Once) [9]，这些方法通过直接在图像上回归实例的位置和类别来进行检测。而 Faster RCNN (Faster Region-based Convolutional Neural Network) [10]作为典型的两阶段方法，其由两部分组成：首先使用区域建议网络生成大量建议框，之后再将建议框中的特征汇集到检测网络中进行检测。此外，两阶段方法还包括 FPN with Faster RCNN (Feature Pyramid Network with Faster Region-based Convolutional Neural Network) [11]、PANet (Path Aggregation Network) [12]、CAFPN (Context-Aware Feature Pyramid Network) [13]等。以上这些方法都在自然图像检测方面取得了显著的突破，但将这些技术直接运用到遥感图像上时却遇到了困难。由于遥感图像是通过卫星或飞机从目标上方获取的，因此普遍具有背景复杂、尺度变化大、目标分布密集、对象容易被遮挡的特点，这给目标检测带来了挑战。

随着卷积神经网络性能的逐步提高，研究人员也提出了多种方法来解决遥感图像中的目标检测问题，目前的主流方法都是通过特征金字塔网络来提取不同大小目标的特征图，并使用路径聚合网络通过横向链接来融合不同层级特征图。在进行边界框回归时使用交并比(Intersection over Union, IoU)作为损失函数也是一种重要的方法，例如广义 IoU 损失(Generalized Intersection over Union, GIoU) [14]、距离 IoU 损失(Distance Intersection over Union, DIoU) [15]和高效 IoU 损失(Efficient Intersection over Union, EIoU) [16]。目前，在遥感图像中的小物体和被遮挡目标检测任务中仍存在如下具有挑战性的问题：1) 小物体在遥感

图像中往往占据不到 20 个像素，这使得在卷积网络中可能会丢失与小物体相关的信息；2) 由于拍摄的视角独特，被检测的目标经常会被云层或树木所遮挡进而导致识别出现漏检或错检；3) 在利用特征金字塔对不同层级特征图进行融合时，不同层级的特征之间也会发生冲突，会降低特征金字塔网络的效果。为了解决以上这些问题，本文提出 IA-YOLO 模型(The network of information enlargement collection and adaptively spatial feature fusion based YOLO)。

2. 本文主要方法

2.1. 模型结构介绍

文章采用 YOLOv5 网络作为基本框架。基于遥感图像小目标图像的特点，本文对网络进行了改进，提出了一种名为 IA-YOLO 的检测网络。其基本框架的组成部分包括 Conv 模块、C3 模块、ConvNeXt 模块[17]、信息扩大集合(Information Enlargement Collection module, IEC)模块、快速空间金字塔池化模块(Spatial Pyramid Pooling Fast, SPPF)、协调注意力模块(Coordinate Attention, CA)、挤压与激励模块(Squeeze and Excitation, SE)和自适应空间特征融合(Adaptively Spatial Feature Fusion, ASFF)模块。Conv 模块对特征图进行卷积、批量归一化(Batch Normalization, BN)和非线性变换。C3 模块采用跨级部分网络的思想，将整个过程的梯度变化融合在特征图中，以减少计算量和模型尺寸。SPPF 通过级联最大池化获得的 3 个尺度特征图来扩展检测模型的感受野。 在检测头部位，添加挤压和激励模块可以减少对冗余信息的计算，提取出与被检测目标相关性更强的通道进行运算。整体结构如图 1 所示。

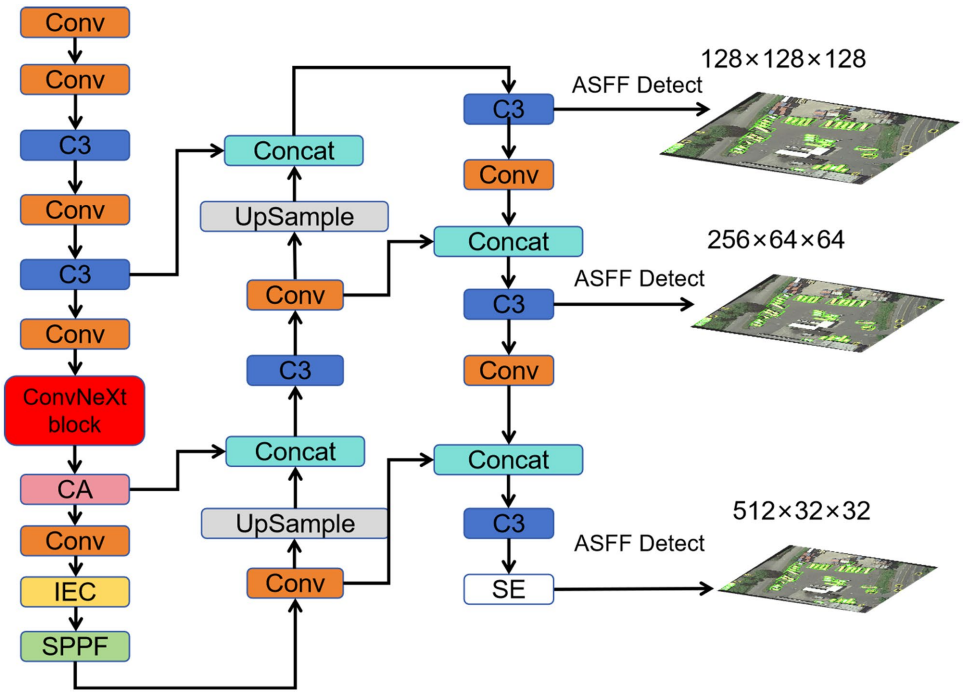

Figure 1. Structure of IA-YOLO
图 1. IA-YOLO 结构图

2.2. 协调注意力模块(Coordinate Attention, CA)

由于遥感数据集当中小目标所占的像素过少，而图像的分辨率相对较大，这就使得其位置信息在融合阶段更容易丢失。为了减少深度网络中微小目标坐标位置信息的丢失率，在 ConvNeXt 模块和下一个

Conv 模块之间添加了协调注意力模块，以改善特征融合之前对不同分辨率的特征图的注意力权重分配。图 2 展示了 Coordinate Attention 模块的注意力编码过程。

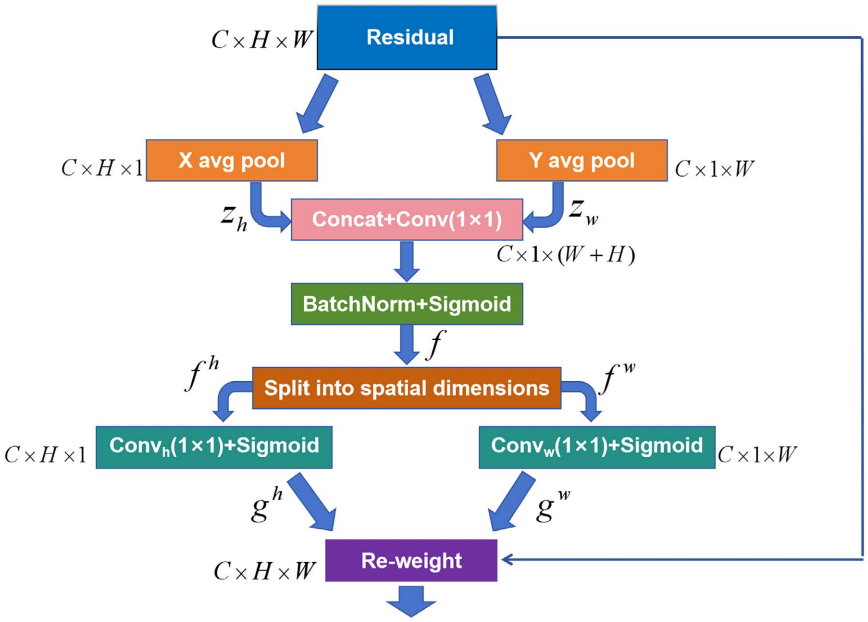

Figure 2. Structure of coordinate attention
图 2. 协调注意力结构图

首先，输入 x_c 是 C 通道、高度为 H、宽度为 W 的特征图。输入是水平坐标方向上的大小$(H, 1)$和垂直坐标方向上的大小$(1, W)$的平均池化。高度为 h 的 c 通道水平编码后的输出表达式为：

$$z_c^h(h)=\frac{1}{W}\sum_{0\le i<W}x_c(h,i) \tag{1}$$

宽度为 w 的通道水平编码后的输出表达式为：

$$z_c^w(w)=\frac{1}{H}\sum_{0\le j<H}x_c(j,w) \tag{2}$$

接下来，对池化编码后的两个通道输出的特征图进行 Concat 拼接，并使用 1 × 1 大小的内核完成级联特征图上的卷积。经过 BatchNorm 层和 Sigmoid 函数后，带有水平和垂直空间信息的输出特征图表示为：

$$f=\delta\left(\mathrm{Conv}_1\left(\left[z^h,z^w\right]\right)\right) \tag{3}$$

将上一步输出的特征图根据不同的空间维度进行划分，得到两个独立的特征图。通过两个 1 × 1 卷积计算并调整和的通道数。它们通过 Sigmoid 激活函数转换为 0 到 1 的范围，作为代表重要性级别的参数。计算公式为：

$$g^h=\sigma\left(\mathrm{Conv}_h\left(f^h\right)\right) \tag{4}$$

$$g^w=\sigma\left(\mathrm{Conv}_w\left(f^w\right)\right) \tag{5}$$

最后，以不同维度的空间信息作为注意力权重对输入 X 进行加权编码，输出 Y 的表达式为：

$$y_c(i,j) = x_c(i,j) \times g_c^h(i) \times g_c^w(j) \tag{6}$$

2.3. ConvNeXt 模块

经过多年的研究以及诸多的优秀论文，“视觉 Transformer 使用全局自注意力特征提取的方法具有优越性”这一点慢慢地被认可。但是文献[18]中提到 Transformer 显示出良好的效果并不是因为其中的全局自注意力模块的优越性。基于此，文献[17]当中提出参考 Swin Transformer 升级 ResNet，进而设计出了 ConvNeXt 模块。

ConvNeXt 模块在 ResNet 的结构基础上吸收了 Transformer 的一些优点。ConvNeXt 模块结构图如图 3 所示。在 ConvNeXt 模块中，使用到了大卷积核的深度卷积，Liu Zhuang [17]等人进行了实验，他们尝试了多种大小的卷积核，最终当核的大小为 7×7 时效果最好。深度卷积完成后的特征图通道数量与输入层的通道数相同，不会增加特征图，但是这种运算对输入层的每个通道独立进行卷积运算，无法有效利用不同通道在相同空间位置上的特征信息，因此需要使用逐点卷积(Pointwise Convolution)来将这些特征图进行组合生成新的特征图。

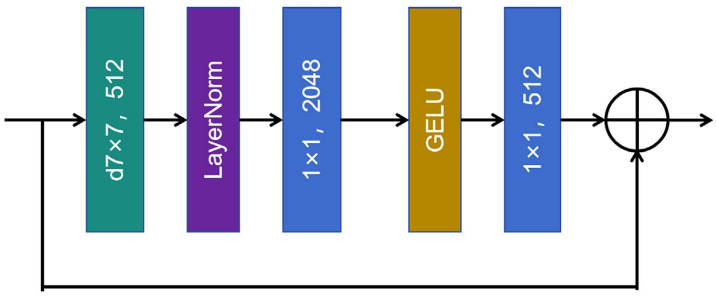

Figure 3. Structrue of ConvNeXt
图 3. ConvNeXt 模块结构

与 ResNet 模块相比，ConvNeXt 模块使用了更少的 Normalization，即仅保留了深度卷积后面的归一化层。在 ResNet 中，每一个 Conv 模块后面都跟随着一个归一化层。ConvNeXt 模块采取了和 Transformer 一样的归一化方法，使得计算更加简单。在 ResNet 模块中所使用的归一化方法是 Batch Normalization (BN)。BN 是神经网络中常用的操作，可以加速网络的收敛并减少过拟合。然而 BN 也有许多复杂的问题，可能会对模型的性能产生负面影响。虽然人们已经开发出了多种标准化技术，但 BN 仍然是大多数视觉任务中的首选。另一方面，Transformers 在网络中使用了更简单的 Layer Normalization (LN)，在不同的应用场景中都获得了良好的性能。因此 ConvNeXt 模块模仿 Tansformer 模块使用 LN 并取得了稍好的性能。另外，ConvNeXt 模块对激活函数的使用方面也学习了 Transformer，Transformer 和 ResNet 模块之间的一个区别是 Transformer 的激活函数较少，而且利用 GELU 函数替换了 ReLU 函数。这种将 ResNet 模块的结构与 Transformer 的风格相结合的做法可以保证 ConvNeXt 模块对图像特征的提取效果在 ResNet 模块之上进一步提高。

2.4. 信息扩大集合模块(Information Enlargement Collection Module)

在对遥感图像进行检测时，有些细长形状的目标只能被检测出一部分甚至被漏检，这是因为模型无法获得目标完整的特征，导致信息出现缺失，难以判断目标边缘，进而使得在框选时出现偏差。因此充分地利用上下文信息对于检测遥感图片中的目标也是非常重要的。近年，使用全局注意或使用更大的卷积核进行卷积操作扩大感受野进而能够获取更多上下文关系的技术得到了广泛地研究，本文提出的信息

扩大集合模块 IEC 便借鉴了这一点，设计结构如图 4 中所示。

IEC 模块是一个完全的卷积网络，其中的上下文信息聚焦块(Contextual Information Focus block，CIF 模块)包含了一个深度卷积和一个空洞深度卷积，卷积核大小均为 7 × 7，空洞卷积的扩张率设置为 3，两个卷积分别构成了两个残差连接。因此，焦点块可以同时与细粒度局部特征和粗粒度全局特征进行交互。尽管取得了一定的效果，但由于输入图像尺寸较大，这些操作通常会带来大量的计算成本和内存开销。

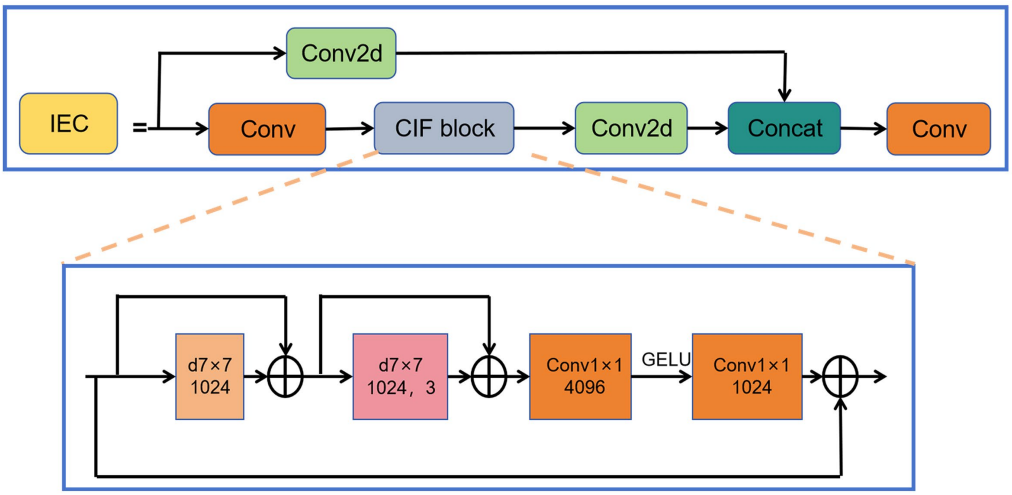

Figure 4. Structure of Information Enlargement Collection module
图 4. 信息扩大集合模块 IEC

2.5. 挤压与激励模块(Squeeze and Excitation, SE)

在 YOLOv5 检测网络的末端，预测头根据先验框对目标进行回归预测。然而，随着特征图的宽与高逐渐减小，通道数会越来越多，会产生许多的冗余信息，这些信息与被检测目标的相关性较低，对提高最终的检测效果没有贡献，只会加大系统计算负担。为了降低这些多余的计算量，需要挑选出与被检测目标相关性较高的信息，着重对包含这些信息的通道进行处理。

挤压与激励模块(Squeeze and Excitation, SE)是一种轻量级注意力机制，可以在保证网络的实时性的同时提升精度，SE 模块分为挤压(Squeeze)和激励(Excitation)两部分。首先通过挤压操作获得全局图像特征；然后，使用激励操作为每个通道生成权重；最后将激励操作的输出权重作为特征通道的重要性指数进行权重分配，对原始的特征图进行加权，重新标定每个通道维度，SE 模块的结构如图 5 所示。在挤压阶段，通过对特征图进行全局平均池化操作，将特征图压缩为 $C \times 1 \times 1$ 的向量，具体计算过程如下：

$$z = F_{sq}(u) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u(i,j) \tag{7}$$

为了利用挤压操作中聚合的信息，接下来进行激励操作，目的是完全捕获通道级别的依赖关系，计算过程可以表示为：

$$\boldsymbol{s} = F_{ex}(z, \boldsymbol{W}) = \sigma(\eta(z, \boldsymbol{W})) = \sigma(\boldsymbol{W}_2 \delta(\boldsymbol{W}_1 z)) \tag{8}$$

在上式当中，δ 代表 ReLU 激活函数，$\boldsymbol{W}_1 \in \boldsymbol{R}^{\frac{C}{r} \times C}$，$\boldsymbol{W}_2 \in \boldsymbol{R}^{\frac{C}{r} \times C}$。

最后将激励操作得到的 $C \times 1 \times 1$ 向量扩展为与输入特征图相同的大小并进行相乘，达到对不同的通道进行权重赋值的目的。

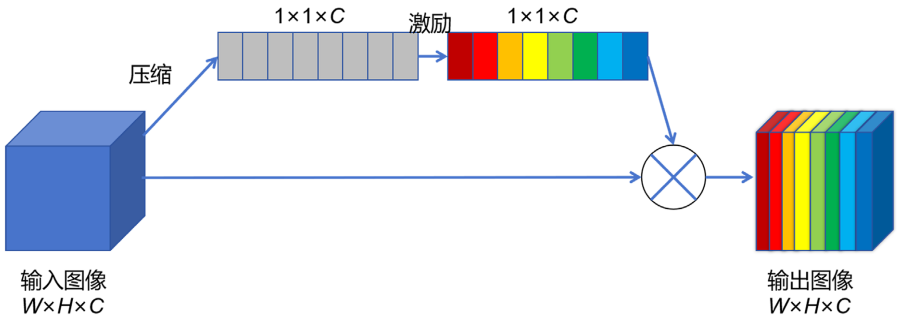

Figure 5. Structure of squeeze and excitation
图 5. 挤压和激励模块

2.6. 自适应空间特征融合模块(Adaptively Spatial Feature Fusion, ASFF)

在 YOLOv5 当中，为了取得不同尺度的特征图，在骨干网络使用了特征金字塔结构。然而，不同特征尺度的特征图之间存在的不一致性会对检测器的效果产生负面影响。在使用特征金字塔检测对象时，特征金字塔上部的特征图中会得到较为抽象的信息，用来负责检测较大的实例目标，而较小的目标的信息往往都保存在底部的特征图当中。在特征金字塔当中，如果目标对象在某个层级的特征图中被分配为正实例，那么在其他层级的特征图中，相应的区域将被视为背景。因此，如果图像同时包含小目标和大目标，那么不同层级的特征图之间就会产生冲突，这些冲突会占据特征金字塔的主要部分，降低特征金字塔的有效性。一些模型采用了几种尝试性策略来处理这个问题。文献[19] [20]将相邻层级特征图的对应区域的梯度设置为零，但这种做法可能会使得相邻层级特征的预测效果变差。文献[21]创建了多个具有不同感受野的分支用于训练和推理。它通过不使用特征金字塔以避免上述问题，也导致其无法使用更高分辨率的特征图，限制了小目标检测的准确性。

本文使用了一种新颖有效的方法，称为自适应空间特征融合(Adaptively Spatial Feature Fusion, ASFF)，来解决特征金字塔所产生的不一致问题，该方法可以对特征金字塔中的特征进行对滤，仅保留有用的信息进行组合，具体结构如图 6 所示。

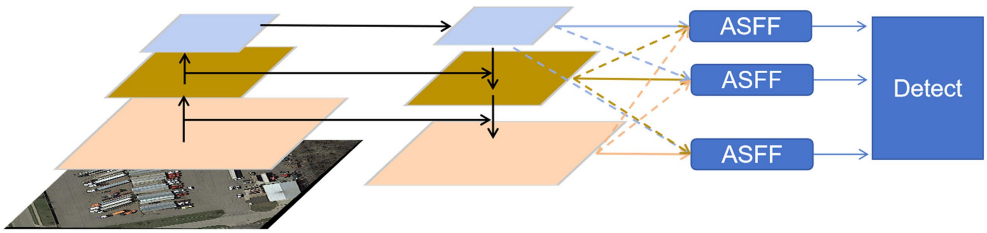

Figure 6. Structure of ASFF
图 6. 自适应空间特征融合模块

对于某一层级的特征，由于 YOLOv5 中的三个层级的特征具有不同的分辨率和通道数，需要先通过上采样和下采样将其他层级的特征进行整合并调整大小到相同的分辨率，对于上采样方法，首先应用 1 × 1 卷积将第 n 层级的通道数改变为与第 l 级相同，再通过插值改变分辨率；对于下采样方法，本文使用步长为 2 的 3 × 3 卷积来同时改变分辨率和通道数。然后再进行自适应融合，融合的计算过程可以表示为下式。

$$y_{ij}^{l} = \alpha_{ij}^{l} \cdot x_{ij}^{1\to l} + \beta_{ij}^{l} \cdot x_{ij}^{2\to l} + \gamma_{ij}^{l} \cdot x_{ij}^{3\to l} \tag{9}$$

其中，$x_{ij}^{n\to l}$ 表示从特征金字塔的第 n 级调整到第 l 级的特征图上位置 (i,j) 处的特征向量。y_{ij}^{l} 表示通道间的输出特征映射到 y^{l} 特征图上 (i,j) 处的特征向量。在 YOLOv5 当中特征金字塔有 3 个层级，故 $n,l\in[1,2,3]$。α_{ij}^{l}，β_{ij}^{l}，γ_{ij}^{l} 是由网络通过自适应学习得到的特征图从第 n 级调整到第 l 级时对应的权重，其中 $\alpha_{ij}^{l}+\beta_{ij}^{l}+\gamma_{ij}^{l}=1$ 且 α_{ij}^{l}，β_{ij}^{l}，$\gamma_{ij}^{l}\in[0,1]$。

3. 实验部分

3.1. 数据集与实验环境

DOTA [22]数据集是遥感领域常用的数据集，具有不同的目标尺度、方向和形状，图片主要来自谷歌地球和中国卫星数据应用中心。图像分辨率从 800 × 800 到 20,000 × 20,000，其中包含 2806 张图片，拥有超过 403,318 个实例目标。DOTA-v1.5 共有 16 个类别，分别为 Plane (PL)、Baseball diamond (BD)、Bridge (BR)、Ground track field (GTF)、Small vehicle (SV)、Large vehicle (LV)、Ship (SH)、Tennis court (TC)、Basketball court (BC)、Storage tank (ST)、Soccer-ball field (SBF)、Roundabout (RA)、Harbor (HA)、Swimming pool (SP)、Helicopter (HC)和 Container Crane (CC)。

MAR-20 数据集[23]是目前规模最大的遥感图像军用飞机目标识别数据集，具有规模大、不同型号目标类间差异小、相同型号目标类内差异大的特点。数据集共计有 3842 张图像，22341 个实例，20 个类别分别记作 A1 至 A20。

NWPU VHR-10 数据集[24]是 NWPU VHR-10 是一个广泛用于遥感图像目标检测的公开数据集，包含 800 张图像，共计 3650 个标注实例，包含密集小目标，总共有 10 个分类，分别为：飞机(Airplane)、舰船(Ship)、储油罐(Storage tank)、棒球场(Baseball diamond)、网球场(Tennis court)、篮球场(Backetball court)、田径场(Ground track field)、港口(Harbor)、桥梁(Bridge)和车辆(Vehicle)。

实验在 Windows 11 系统下运行，所有使用的深度学习模型均使用 torch2.0.1 版本的 Pytorch 框架实现，训练使用的 GPU 显卡为 NVIDIA GeForce RTX 4060，所用显存为 16GB，Cuda 版本为 11.8。CPU 为 Intel(R)Core(TM) i9-13900H。初始学习率设置为 0.01，训练迭代 200 轮，batchsize 为 8，并选择 AP 和 mAP 作为评价指标进行测试。

3.2. 评价指标

实验采用的评价指标为精确率(Precision, P)、召回率(Recall, R)、平均精度均值(mean Average Prescision, mAP)、平均精确率(Average Prescision, AP)作为评估指标，计算公式如下：

$$P=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}\times 100\% \tag{10}$$

$$R=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}\times 100\% \tag{11}$$

$$\mathrm{AP}=\int_0^1 P(R)\,\mathrm{d}R\times 100\% \tag{12}$$

$$\mathrm{mAP}=\frac{1}{n}\sum_{i=1}^{n}\mathrm{AP}_i\times 100\% \tag{13}$$

其中，TP 为将正确目标预测为正样本，FP 为错误将非目标预测为正样本的数量，FN 为漏检实际目标的数量，n 为检测对象类别，AP 为每一种类别的平均精度值，mAP 为所有类别的平均精度值。

3.3. 对比试验

为了验证本文方法的有效性，将本文方法与当前在遥感目标检测领域先进的方法进行对比，表 1~3

分别是在 3 个不同数据集上的性能比较。

由表中数据可见，本文的方法在多个数据集上都表现出了出色的性能。其中，在 DOTA 数据集的 16 类别当中，本文方法取得了 8 个类别最高的 mAP，与经典的 Faster R-CNN 相比，本文方法的精度提升了 15.3 个百分点，特别是 SV 和 SH 两个类别，分别提升了 40.7 个百分点和 54.7 个百分点，体现出了本文方法在检测小像素目标方面的优良性能。在 MAR20 数据集当中，本文方法的精度相较于 Faster R-CNN 提升了 7 个百分点，说明本文提出的模型能够很好地区分外形相似的目标，这得益于本文模型良好的特征提取能力，证明了模型骨干网络的性能。在 NWPU VHR10 数据集当中，本文方法比 Mask R-CNN 高出了 15.2 个百分点。harbor 类别的提升尤为突出，该类别目标长宽比大，形状狭长，这说明本文方法极大改善了长宽比较大的目标的检测性能，而其他算法难以契合该特点的目标，导致结果不佳。

为了进一步证明本文方法的有效性，实验对 3 个数据集进行了可视化，效果如图 7~9 所示。在图 7 中可以发现，当尺寸差异较大的球场环岛和小型车辆同时出现时，也能精准地检测出小目标且几乎没有漏检，说明本文的方法成功地提取到了小目标的特征且在多尺度融合阶段没有造成缺失。在图 8 中，面对形状基本相似的战机，仅凭借有无鸭翼这种微小的外观差异便能成功进行分类，说明本文的方法能够更全面地捕获目标的相关信息，避免了由于缺少部分特征信息而导致的错检问题。在图 9 中，有许多的船类小目标，它们密集地排列在形状狭长的港口目标周围，本文的方法能够同时将这些目标全部检测出来，这是因为本文的模型可以充分地利用每个像素点的上下文信息来找到目标的边界轮廓。通过使用更加大范围的卷积核，即便是狭长形状的港口，也可以完整地得到全部的边界位置的特征信息。

Table 1. Performance comparison with different algorithms on MAR20
表 1. 不同算法在 MAR20 数据集上的对比

类别	RetinaNet	FCOS	ATSS	Faster R-CNN	IA-YOLO
A1	70.7	71.1	69.9	79.7	81.6
A2	82.1	79.4	83.7	81.6	93.9
A3	45.7	67.1	70.3	84.7	92.3
A4	46.1	46.4	50.2	72.0	91.1
A5	70.4	65.1	59.1	75.0	76.4
A6	76.7	84.8	85.1	88.0	96.9
A7	50.7	72.1	82.6	87.1	94.8
A8	81.2	83.3	80.0	89.7	91.0
A9	50.8	80.5	84.4	89.6	88.8
A10	88.9	84.5	90.0	90.6	97.2
A11	36.2	57.8	61.4	82.5	85.7
A12	77.0	71.3	82.7	86.4	87.2
A13	62.7	58.5	68.1	66.6	82.7
A14	71.8	75.3	78.1	85.8	92.1
A15	25.6	33.8	40.5	45.8	59.3
A16	85.0	83.7	87.0	87.8	91.8
A17	85.5	87.1	89.5	90.2	95.3
A18	27.9	38.6	36.1	54.2	67.4
A19	68.1	69.2	67.7	76.6	80.3
A20	77.7	66.4	77.5	78.6	85.8
mAP (%)	64.0	68.8	72.2	79.6	86.6

Table 2. Performance comparison with different algorithms on DOTA
表 2. 不同算法在 DOTA 数据集上的对比

检测方法	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	CC	mAP (%)
SSD	78.1	38.6	14.5	27.4	13.5	39.5	30.6	81.7	42.0	43.7	20.7	33.6	46.9	39.3	30.4	0	36.3
FasterRCNN	70.0	63.4	55.5	63.3	22.7	44.5	32.5	81.3	68.4	36.9	57.6	49.8	59.1	55.1	51.5	16.9	51.8
YOLOv4	85.2	61.7	32.6	35.2	37.0	64.0	79.5	88.3	55.6	64.8	34.4	54.2	69.8	58.5	67.6	0.7	55.6
FPN	78.6	70.1	55.1	68.5	23.7	45.4	40.3	86.0	69.4	46.4	61.1	56.2	59.5	64.5	68.3	24.4	57.3
PANet	85.9	74.1	51.5	64.5	27.7	56.2	58.4	89.6	67.0	61.3	63.4	59.2	67.9	**73.4**	**71.3**	7.6	61.2
CDDNet	81.4	74.7	55.3	70.1	25.3	51.5	49.2	89.8	71.4	53.3	65.6	58.2	69.9	60.4	**71.3**	32.7	61.3
HCENet	79.4	**76.0**	**64.7**	**74.5**	24.4	48.6	43.4	89.8	**74.0**	50.0	**66.0**	63.2	70.1	66.6	68.7	**33.1**	62.0
IA-YOLO	**90.6**	72.9	49.7	55.7	**63.4**	**81.0**	**87.2**	**93.9**	63.9	**79.8**	47.0	**64.4**	**77.2**	71.7	62.0	13.0	**67.1**

Table 3. Performance comparison with different algorithms on NWPU VHR10
表 3. 不同算法在 NWPU VHR10 数据集上的对比

类别	Fast R-CNN	Mask R-CNN	RetinaNet	FCOS	CenterNet	SGFTHR	IA-YOLO
Airplane	78.0	80.8	78.1	77.8	75.9	78.1	99.1
Ship	66.3	67.0	65.6	66.8	65.7	67.1	85.5
Storage tank	84.8	84.2	84.8	85.0	83.1	86.4	91.6
Tennis court	76.5	87.7	66.8	81.7	72.7	85.8	78.1
Basketball court	43.3	45.7	44.8	40.1	43.8	46.1	60.6
Ground track Field	86.6	85.7	88.7	87.1	86.9	87.4	99.5
Harbor	36.6	36.8	35.6	37.5	35.4	37.9	90.5
Bridge	49.0	47.6	52.6	49.5	48.5	50.9	72.2
Vehicle	79.7	76.9	68.8	87.3	72.8	93.3	86.7
Baseball Diamond	95.1	97.0	96.8	94.1	92.5	97.4	60.6
mAP (%)	69.6	70.9	68.3	70.7	67.7	73.1	86.1

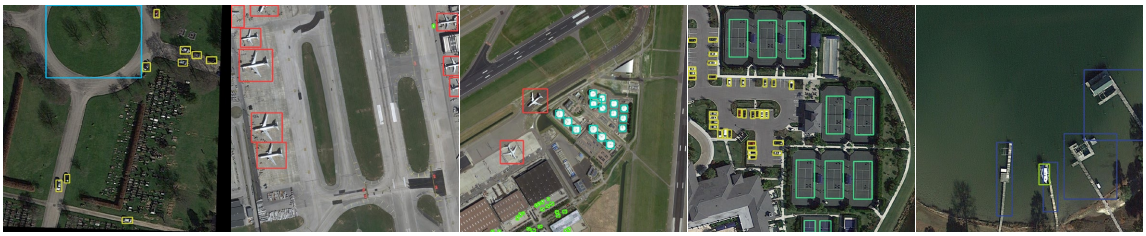

Figure 7. Visualization of detection results on the DOTA
图 7. DOTA 数据集可视化结果

Figure 8. Visualization of detection results on the MAR20
图 8. MAR20 数据集可视化结果

Figure 9. Visualization of detection results on the NWPU VHR10
图 9. NWPU VHR10 数据集可视化结果

3.4. 消融实验

为了验证每个模块的贡献，文章中分别使用基线模型和各个模块进行组合，在 DOTA 数据集上进行消融实验，实验结果如表 4 所示，本文提出的方法比基线的 mAP@0.5 提高了 2.5 个百分点。CA 模块对不同分辨率的特征图的注意力权重分配，提取出关注全局空间特征的重要部分，通过强化对目标边界处的特征提取，提升了 12 个类别目标的 mAP 值，说明了此模块的泛用性。在基线当中加入 ConvNeXt 模块后类别 BD 的 mAP 提升了 4.3 个百分点，类别 HC 提升了 6.6 个百分点，说明此模块对边界颜色与环境背景颜色有一定区分度的目标较为敏感。通过使用 IEC 模块，模型获取到了更多的周边像素信息，将类别 HA 的 mAP 值提升了 1.5 个百分点，说明该模块使得模型更利于检测长宽比较大的目标。对类别 SP 这种边界轮廓容易遭到遮挡的目标，IEC 模块可以更好地利用上下文信息判断出目标的边界，并将类别 SP 的 mAP 值提高到了 72.9%。在 SE 模块的作用下，类别 CC 和类别 HC 的 mAP 分别提升了 6.52 个百分点和 9.3 个百分点。与以上 4 个模块相比，基线添加 ASFF 模块后，有 7 个类别取得了最高的 mAP。

Table 4. Module ablation experiment
表 4. 模块消融实验

检测方法	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	CC	mAP (%)
Baseline	90.0	69.4	45.6	49.3	62.2	81.1	87.0	93.0	62.4	78.1	49.4	63.7	77.2	71.5	52.9	0.36	64.6
+CA	90.5	72.3	48.4	54.2	62.5	80.6	87.0	93.4	63.7	79.0	53.3	65.9	77.9	**72.9**	56.8	0.59	66.2
+SE	90.6	72.5	49	53.7	62.3	80.7	87.2	93.6	63.5	78.7	55.0	65.5	78.2	68.9	62.2	6.88	66.8
+ConvNeXt	89.9	**73.7**	48.6	50.6	62.4	80.5	86.7	93.7	62.2	78.1	50.0	63.6	77.0	70.9	59.5	0.37	65.5
+IEC	90.4	72.1	46.9	54.7	62.4	80.8	**87.3**	93.6	62.6	78.8	**55.0**	61.3	**78.7**	69.9	54.5	6.09	66.0
+ASFF_Detect	**90.9**	71.3	48.0	52.8	63.1	**81.7**	**87.3**	93.6	**64.3**	**80.6**	50.3	**66.2**	77.5	70.4	**65.0**	1.48	66.5
IA-YOLO	90.6	72.9	**49.7**	**55.7**	**63.4**	81.0	87.2	**93.9**	63.9	79.8	47.0	64.4	77.2	71.7	62.0	**13.0**	67.1

3.5. 检测结果分析

图 10 展示了在 DOTA 数据集下本文方法与基线方法的检测结果比较，其中第一列为真实框，第二列为基线方法的结果，第三列为本文的结果。通过第一行可以发现，相较于原图的真实框，基线方法会出现错检和漏检情况，将右下方的方形阴影判别为了小型车辆。在中间位置，由于目标排列比较紧密且形状相似，发生了漏检。本文提出的方法并没有出现这些情况。在第二行当中图片下方的车辆位于树木的遮挡之下，基线方法没有能够将其标注出来，本文的方法能准确地将其识别出来。

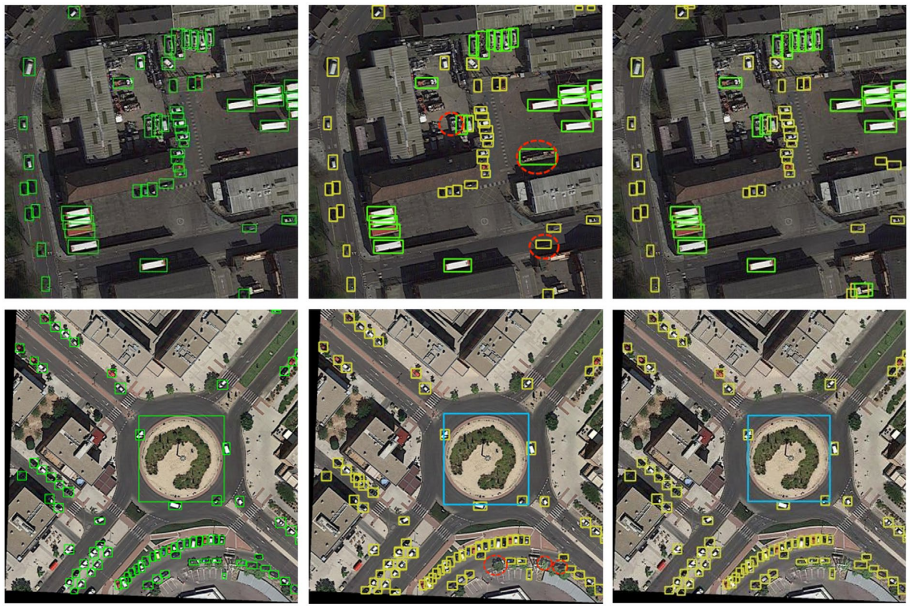

Figure 10. Comparison chart of test results (a) Bounding box; (b) Baseline; (c) IA-YOLO
图 10. 检测结果对比图 (a) 真实框；(b) 基线方法；(c) IA-YOLO

改进前后模型的平均精度均值(mAP)曲线如图 11 所示，图 12 展现了改进前后模型的 PR 曲线图，图中最粗的一根曲线与两条轴所围成的区域的大小就代表了 mAP 的大小，可以看出 IA-YOLO 的效果更佳，进一步说明了本文方法的有效性。表 5 展示了本文方法与基线方法的性能差距。虽然参数量和计算量都有所增加，但是 mAP 提升了 2.5 个百分点，而且在对 1024 × 1024 大小的图像进行推理时所消耗的时间仅增加了 4.4 毫秒。

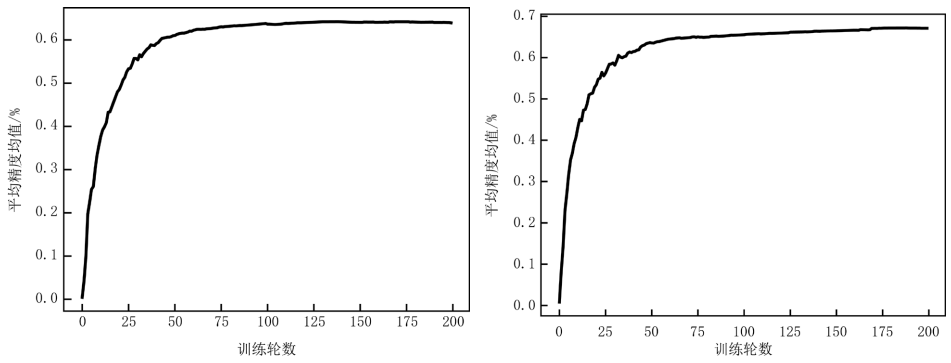

Figure 11. Performance of the model before and after improvement
图 11. 改进前模型的性能

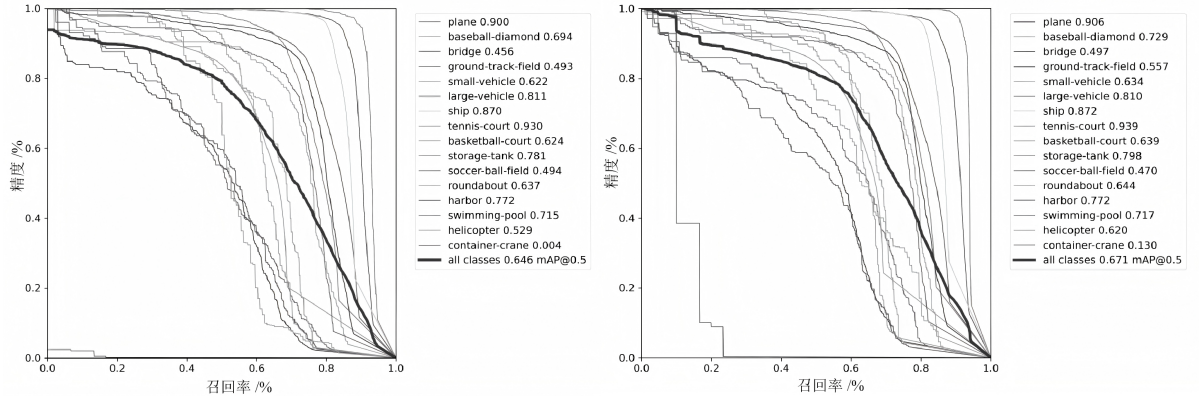

Figure 12. PR curve of the model before and after improvement
图 12. 改进前后模型的 PR 曲线

Table 5. Comparison of model performance before and after improvements
表 5. 改进前后模型性能比较

使用方法	mAP	GFLOPs	Para	推理时间
Yolov5	64.6%	16.0	7.1 M	11.6 ms
IA-YOLO	67.1%	24.7	12.3 M	15.9 ms

4. 结论

本文基于 YOLOv5 网络，提出了一种用于遥感下目标检测的 IA-YOLO 网络，在 DOTA 数据集上，该网络的 mAP@0.50 达到了 67.1%，相较于原始 YOLOv5 模型，提升了 2.5 个百分点。该模型的主要贡献如下：1) 在骨干网络中，使用 ConvNeXt 模块可以更好地获得图片中的特征并减少过拟合，当中的大卷积核可以扩大特征提取的范围、使得网络的分辨能力更强，以此来确定被遮挡目标的位置和类别。2) 由于遥感数据集当中小目标所占的像素过少，而相对来讲，图像的分辨率要大得多，这就使得其位置信息在融合阶段更容易丢失，添加协调注意力(CA)可以帮助网络注意到小目标，防止其位置信息丢失。3) 对于长宽比较大的目标，本文提出的扩大信息集合模块(IEC)通过扩大卷积核中元素的间隔，使得间距更远的图像像素的特征信息联系在一起，充分利用遥感图像包含的丰富的上下文语义信息，使得模型可以更加全面地寻找目标的边界。4) 遥感图像中大目标与小目标的大小差距过大，在多尺度融合阶段必然会造成冲突，为了保证特征金字塔的效果，缓解小目标特征被覆盖的现状，本文采用了自适应空间特征融合模块(ASFF)来消除不同层级特征图之间的冲突，防止小目标的特征信息被大目标信息淹没。5) 使用挤压与激励模块(SE)突出前景检测目标，弱化背景环境的同时，减少了模型的参数量。

5. 展望

本文基于 YOLOv5 提出了一种新型的小目标检测器 IA-YOLO 并取得了一定的成果，未来的研究方向可以基于小目标的方向任意性，在确保检测效果和运行速度的前提下，设计利用旋转框的对目标进行框选的检测器，使用旋转框可以更加显著地描绘出图像当中目标的大小和形状，在实际生活中可以有更加广泛的应用。此外为了保证模型的提取小目标信息的能力，可以设计全新的骨干网络，并在预处理阶段使用多样的数据增强方式，增添训练图像的多样性。

基金项目

国家自然科学基金资助项目(61603255)资助。

参考文献

[1] 陈磊, 张孙杰, 王永雄. 基于改进的 YOLOv3 及其在遥感图像中的检测[J]. 小型微型计算机系统, 2020, 41(11): 2321-2324.

[2] Wang, Q., Gao, J. and Li, X. (2019) Weakly Supervised Adversarial Domain Adaptation for Semantic Segmentation in Urban Scenes. *IEEE Transactions on Image Processing*, **28**, 4376-4386. https://doi.org/10.1109/tip.2019.2910667

[3] Jin, R. and Lin, D. (2020) Adaptive Anchor for Fast Object Detection in Aerial Image. *IEEE Geoscience and Remote Sensing Letters*, **17**, 839-843. https://doi.org/10.1109/lgrs.2019.2936173

[4] Gerhards, M., Schlerf, M., Mallick, K. and Udelhoven, T. (2019) Challenges and Future Perspectives of Multi-/Hyperspectral Thermal Infrared Remote Sensing for Crop Water-Stress Detection: A Review. *Remote Sensing*, **11**, Article 1240. https://doi.org/10.3390/rs11101240

[5] Li, X., Chen, M., Nie, F., *et al*. (2017) A Multiview-Based Parameter Free Framework for Group Detection. *AAAI Conference on Artificial Intelligence*, San Francisco, 4-9 February 2017, 4147-4153.

[6] Zheng, Q., Ma, J., Liu, M., Liu, Y., Li, Y. and Shi, G. (2022) Lightweight Hot-Spot Fault Detection Model of Photovoltaic Panels in UAV Remote-Sensing Image. *Sensors*, **22**, Article 4617. https://doi.org/10.3390/s22124617

[7] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C., *et al*. (2016) SSD: Single Shot Multibox Detector. *Computer Vision—ECCV* 2016, Amsterdam, 11-14 October 2016, 21-37. https://doi.org/10.1007/978-3-319-46448-0_2

[8] Lin, T., Goyal, P., Girshick, R., He, K. and Dollar, P. (2017) Focal Loss for Dense Object Detection. 2017 *IEEE International Conference on Computer Vision* (*ICCV*), Venice, 22-29 October 2017, 2999-3007. https://doi.org/10.1109/iccv.2017.324

[9] Redmon, J., Divvala, S., Girshick, R. and Farhadi, A. (2016) You Only Look Once: Unified, Real-Time Object Detection. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 779-788. https://doi.org/10.1109/cvpr.2016.91

[10] Ren, S., He, K., Girshick, R. and Sun, J. (2017) Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **39**, 1137-1149. https://doi.org/10.1109/tpami.2016.2577031

[11] Lin, T., Dollar, P., Girshick, R., He, K., Hariharan, B. and Belongie, S. (2017) Feature Pyramid Networks for Object Detection. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 936-944. https://doi.org/10.1109/cvpr.2017.106

[12] Liu, S., Qi, L., Qin, H., Shi, J. and Jia, J. (2018) Path Aggregation Network for Instance Segmentation. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 8759-8768. https://doi.org/10.1109/cvpr.2018.00913

[13] 熊娟, 张孙杰, 阚亚亚, 等. 基于 CAFPN 和细化双头解耦的遥感图像目标检测[J]. 应用科学学报, 2023, 41(6): 989-1003.

[14] Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I. and Savarese, S. (2019) Generalized Intersection over Union: A Metric and a Loss for Bounding Box Regression. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 658-666. https://doi.org/10.1109/cvpr.2019.00075

[15] Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R. and Ren, D. (2020) Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression. *Proceedings of the AAAI Conference on Artificial Intelligence*, **34**, 12993-13000. https://doi.org/10.1609/aaai.v34i07.6999

[16] Zhang, Y., Ren, W., Zhang, Z., Jia, Z., Wang, L. and Tan, T. (2022) Focal and Efficient IoU Loss for Accurate Bounding Box Regression. *Neurocomputing*, **506**, 146-157. https://doi.org/10.1016/j.neucom.2022.07.042

[17] Liu, Z., Mao, H., Wu, C., Feichtenhofer, C., Darrell, T. and Xie, S. (2022) A Convnet for the 2020s. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, 18-24 June 2022, 11966-11976. https://doi.org/10.1109/cvpr52688.2022.01167

[18] Tolstikhin, I.O., Houlsby, N., Kolesnikov, A., *et al* (2021) MLP-Mixer: An All-MLP Architecture for Vision. *Advances in Neural Information Processing Systems*, **34**, 24261-24272.

[19] Wang, J., Chen, K., Yang, S., Loy, C.C. and Lin, D. (2019) Region Proposal by Guided Anchoring. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 2960-2969. https://doi.org/10.1109/cvpr.2019.00308

[20] Zhu, C., He, Y. and Savvides, M. (2019) Feature Selective Anchor-Free Module for Single-Shot Object Detection. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 840-849. https://doi.org/10.1109/cvpr.2019.00093

[21] Li, Y., Chen, Y., Wang, N. and Zhang, Z. (2019) Scale-Aware Trident Networks for Object Detection. 2019 *IEEE/CVF*

International Conference on Computer Vision (*ICCV*), Seoul, 27 October-2 November 2019, 6053-6062. https://doi.org/10.1109/iccv.2019.00615

[22] Xia, G., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., *et al*. (2018) DOTA: A Large-Scale Dataset for Object Detection in Aerial Images. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 3974-3983. https://doi.org/10.1109/cvpr.2018.00418

[23] 禹文奇, 程塨, 王美君, 等. MAR20: 遥感图像军用飞机目标识别数据集[J]. 遥感学报, 2023, 27(12): 2688-2696.

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