智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

改进YOLOv5s的遥感图像目标检测

赵文清, 康怿瑾, 赵振兵, 翟永杰

引用本文:
赵文清,康怿瑾,赵振兵,翟永杰. 改进YOLOv5s的遥感图像目标检测[J]. 智能系统学报, 2023, 18(1): 86–95.
ZHAO Wenqing,KANG Yijin,ZHAO Zhenbing,ZHAI Yongjie. A remote sensing image object detection algorithm with improved YOLOv5s[J]. *CAAI Transactions on Intelligent Systems*, 2023, 18(1): 86–95.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202203013

您可能感兴趣的其他文章

自适应上下文特征的多尺度目标检测算法
Multi-scale target detection algorithm based on adaptive context features
智能系统学报. 2022, 17(2): 276–285 https://dx.doi.org/10.11992/tis.202101029

多尺度特征融合网络的视网膜OCT图像分类
Retinal optical coherence tomography image classification based on multiscale feature fusion
智能系统学报. 2022, 17(2): 360–367 https://dx.doi.org/10.11992/tis.202111024

一种轻量化油田危险区域入侵检测算法
A lightweight intrusion detection algorithm for hazardous areas in oilfields
智能系统学报. 2022, 17(3): 634–642 https://dx.doi.org/10.11992/tis.202107033

双向特征融合与注意力机制结合的目标检测
Target detection based on bidirectional feature fusion and an attention mechanism
智能系统学报. 2021, 16(6): 1098–1105 https://dx.doi.org/10.11992/tis.202012029

基于跳跃连接金字塔模型的小目标检测
Skip feature pyramid network with a global receptive field for small object detection
智能系统学报. 2019, 14(6): 1144–1151 https://dx.doi.org/10.11992/tis.201905041

DOI: 10.11992/tis.202203013
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.TP.20220930.1620.004.html.

改进 YOLOv5s 的遥感图像目标检测

赵文清[1,2], 康怿瑾[1], 赵振兵[3], 翟永杰[1]

（1. 华北电力大学 控制与计算机工程学院, 河北 保定 071003; 2. 复杂能源系统智能计算教育部工程研究中心, 河北 保定 071003; 3. 华北电力大学 电气与电子工程学院, 河北 保定 071003）

摘　要: 针对遥感图像中感兴趣目标特征不明显、背景信息复杂、小目标居多导致的目标检测精度较低的问题, 本文提出了一种改进 YOLOv5s 的遥感图像目标检测算法（Swin-YOLOv5s）。首先, 在骨干特征提取网络的卷积块中加入轻量级通道注意力结构, 抑制无关信息的干扰; 其次, 在多尺度特征融合的基础上进行跨尺度连接和上下文信息加权操作来加强待检测目标的特征提取, 将融合后的特征图组成新的特征金字塔; 最后, 在特征融合的过程中引入 Swin Transformer 网络结构和坐标注意力机制, 进一步增强小目标的语义信息和全局感知能力。将本文提出的算法在 DOTA 数据集和 RSOD 数据集上进行消融实验,结果表明,本文提出的算法能够明显提高遥感图像目标检测的平均准确率。

关键词: 遥感图像; 感兴趣目标; 目标检测; 特征提取; 轻量级通道注意力结构; 多尺度特征融合; 上下文信息; Swin 变换器; 坐标注意力机制
中图分类号: TP751;TP391　　**文献标志码**: A　　**文章编号**: 1673−4785(2023)01−0086−10

中文引用格式: 赵文清, 康怿瑾, 赵振兵, 等. 改进 YOLOv5s 的遥感图像目标检测 [J]. 智能系统学报, 2023, 18(1): 86–95.
英文引用格式: ZHAO Wenqing, KANG Yijin, ZHAO Zhenbing, et al. A remote sensing image object detection algorithm with improved YOLOv5s[J]. CAAI transactions on intelligent systems, 2023, 18(1): 86–95.

A remote sensing image object detection algorithm with improved YOLOv5s

ZHAO Wenqing[1,2], KANG Yijin[1], ZHAO Zhenbing[3], ZHAI Yongjie[1]

(1. School of Control and Computer Engineering, North China Electric Power University, Baoding 071003, China; 2. Engineering Research Center of the Ministry of Education for Intelligent Computing of Complex Energy System, Baoding 071003, China; 3. School of Electrical and Electronic Engineering, North China Electric Power University, Baoding 071003, China)

Abstract: Aiming at the low average target detection accuracy in remote sensing images caused by obscure features in the objects of interest, complex background information, and multiple small targets, we propose a new remote sensing image object detection algorithm with improved YOLOv5s (Swin-YOLOv5s). First, an efficient channel attention structure is added to the convolutional block of the backbone feature extraction network to suppress the interference of irrelevant information; second, cross-scale connection and contextual information weighting operations are performed to enhance detection target feature extraction on the basis of multiscale feature fusion, and the fused feature maps are composed into a new feature pyramid; finally, the Swin Transformer structure and coordinate attention mechanism are used to further enhance the semantic information and global perception ability of small targets. The result of a feature fusion elimination experiment performed on the DOTA and RSOD datasets shows that the proposed algorithm can significantly improve the average accuracy of object detection in remote sensing images.

Keywords: remote sensing images; objects of interest; object detection; feature extraction; efficient channel attention structure; multiscale feature fusion; contextual information; Swin Transformer; coordinate attention mechanism

遥感图像目标检测广泛应用于军事国防、海洋检测、智能交通、突发灾害和应急响应等各个方面。遥感图像目标检测旨在从复杂的遥感背景图像中找到感兴趣的目标, 并精确高效的标注其位置和类别。与现有的自然图像相比, 遥感图像背景信息复杂、待检测目标的方向具有不确定

收稿日期: 2022−03−08.　**网络出版日期**: 2022−10−06.
基金项目: 河北省自然科学基金项目 (F2021502013); 中央高校基本科研业务费面上项目 (2020MS153, 2021PT018); 国家自然科学基金项目 (61773160, 61871182).
通信作者: 赵文清. E-mail: jbzwq@126.com.

©《智能系统学报》编辑部版权所有

性、尺度变化大。遥感图像经过多次卷积池化操作之后到达网络深层的小目标信息会逐渐丢失, 导致目标检测平均准确率降低。

现阶段, 主流的目标检测算法分为两阶段: 目标检测算法和单阶段目标检测算法。Girshick 等陆续提出了 R-CNN (region-based convolutional neural network)[1]、Fast R-CNN[2]、Faster R-CNN[3] 等两阶段目标检测算法, 检测精度高但速度慢。于是, Redmon 等提出了 YOLO[4]、YOLOv2[5]、YOLOv3[6] 等单阶段算法, 舍弃了候选框生成阶段, 直接对目标进行分类和回归操作, 提高了目标检测算法的实时检测速度。Bochkovskiy 等提出了 YOLOv4[7], 在 FPN[8] 的基础上添加了一条自底向上的路径, 并使用路径聚合网络 (PANet)[9] 来提高语义信息流在网络中的特征传递效率。Glenn 等提出了 YOLOv5, 在 YOLOv4[7] 的基础上增加了一些训练技巧, 最为突出的是在骨干网络中引入了 Focus 结构, 减少模型参数量, 提高模型利用率。

林娜等[10] 运用空洞残差卷积的思想提取浅层特征, 将提取到的特征与深层特征进行融合, 有效提高了遥感图像中飞机的检测精度。姚艳清等[11] 使用双尺度特征融合模块, 缓解深层信息的丢失问题, 有效提高了多尺度遥感目标的检测能力。张晓雅等[12] 提出了多阶段级联结构的遥感图像目标检测算法, 在水平框和旋转框两个检测任务上均有提升。以上这些方法都是基于两阶段目标检测算法, 通过增加感受野并结合浅层特征与深层特征进行多尺度特征融合的方式, 在一定程度上提高了目标检测的能力, 但是实时性太差。李婕等[13] 提出了基于平行层特征共享的遥感检测模型 AFF-CenterNet, 显著提高了飞机中小目标的表征能力。虽然保持了单阶段目标检测算法的速度优势, 但是其研究类别单一, 无法推广到多类别遥感目标检测应用中。

现有的通用目标检测算法大都是基于卷积神经网络进行的改进, 虽然卷积神经网络能够有效提取局部信息, 但是卷积运算的局部性限制了它获取全局上下文信息的能力。一些学者陆续探索了新的方法, 如: Zhang 等[14] 基于 YOLOv4-P7 模型提出了一种融合卷积和自注意力模块的混合模型 ViT-YOLO, 使用加权双向特征金字塔网络 (Bi-directional feature pyramid network, BiFPN)[15] 用于跨尺度特征融合, 并由原来的 3 个检测头变为 5 个检测头, 在 VisDrone2019 数据集中有效提高了小目标的平均检测精度值, 由原来的 35.4% 提高到 38.5%, 但是增加了计算量和存储成本。Zhu 等[16] 提出了 TPH-YOLOv5 模型, 在 YOLOv5 的基础上增加了具有 Transformer 特性的检测头和 CBAM[17] 模块, 在 VisDrone2021 数据集上有效提高了小目标的平均检测精度值。Xu 等[18] 提出了基于 Swin Transformer[19] 的局部感知骨干网络 (local perception swin transformer, LPSW), 进行遥感图像目标检测与实例分割的探索, 以增强网络的局部感知能力, 提高小尺度物体的检测精度。

以上方法均以 Transformer 结构为载体, 通过多头自注意力模块为目标检测保留足够的空间信息, 在无人机航拍图像数据集上进行了实验验证, 实验结果表明该结构能在一定程度上提高小目标的检测性能。但是仍然存在以下问题: 1)对小尺度目标检测的性能较低、局部信息获取能力较弱; 2)当网络相对较浅且特征图分辨率相对较大时, Transformer 结构被过早地用于特征提取时, 这可能会丢失一些待检测目标的上下文信息; 3)Transformer 在预测密集目标的场景下, 对方向任意、尺度变化范围大、大中小目标分布不均匀的遥感图像的目标检测具有很大的影响, 其计算复杂度过大, 导致实时性变差。故本文结合 Transformer 的思想, 引入了 Swin Transformer 网络, 在多尺度特征融合的过程中将其嵌入到卷积结构, 通过多头自注意力模块来增强遥感图像中小目标的语义信息和特征表示, 减少计算量。

综上所述, 本文结合遥感图像中背景信息复杂、小目标检测困难、语义信息丢失严重的问题, 提出了改进 YOLOv5s 的遥感图像目标检测算法 (Swin-YOLOv5s)。首先, 针对遥感图像背景信息复杂的问题, 在骨干网络的卷积块中加入了一种即插即用的轻量级通道注意力结构 (efficient channel attention, ECA)[20], 形成新的卷积块 (convolutional efficient channel attention, CECA), 使用不降维的局部跨信道交互策略加强遥感目标的特征提取能力。其次, 在 FPN[9] 的基础上引入一种更为高效简洁的 BiFPN[15] 多尺度特征融合网络, 进行有效的跨尺度连接和上下文信息加权操作, 避免了大量小目标语义信息的丢失。最后, 在多尺度特征融合的过程中引入具有 Swin Transformer[18] 网络特性的 C3STR (cross stage partial bottleneck with 3 convolutions and swin transformer) 模块和坐标注意力机制, 以增强网络的局部感知能力, 提高小尺度目标的检测精度。

1　相关技术和理论

YOLOv5s 目标检测算法主要由输入端、骨干网络、特征融合、目标位置与类别预测层四部分

组成，其中训练图像输入尺寸设置为 640×640。由于 DOTA 数据集小目标分布均匀、中大目标严重分布不均，在数据预处理阶段使用 Mosaic 数据增强随机选取 4 张图片进行拼接，能极大程度上改变目标分布不均匀的情况，并且随着训练时间的加长，改善效果越明显。

YOLOv5s 以 CSPDarknet 为骨干网络，主要包括 Focus 切片结构、跨阶段局部网络 (cross stage partial networks, CSPNet)、空间金字塔池化 (spatial pyramid pooling, SPP) 模块等；特征融合部分继续沿用了 YOLOv4 的自顶向下和自底向上的多尺度特征融合方式，然后将提取到的特征传入到检测层，输出尺度分别为 80×80、40×40、20×20 来实现大中小目标的类别和位置预测，最后通过非极大值抑制算法 (non maximum suppression, NMS) 等后处理操作消除冗余框，输出置信度得分最高的预测物体的类别。

2　改进 YOLOv5s 的遥感图像目标检测

本文提出的改进 YOLOv5s 的遥感图像目标检测算法（Swin-YOLOv5s）的整体框架结构图如图 1 所示。在骨干网络的卷积块中加入轻量级通道注意力结构形成新的卷积块 CECA，进而抑制复杂背景信息，增强小目标信息提取能力。由于遥感图像中目标比较密集，容易出现漏检和误检情况，因此在多尺度特征融合的基础上，进行跨尺度连接和上下文信息加权特征融合操作，在不增加计算成本的情况下融合更多不同尺度的特征。使用 C3STR 结构克服 CNN 卷积操作的局限性,识别底层特征的抽象信息，增强局部几何特征信息的感知能力；通过对融合后的特征图使用坐标注意力机制进行更新。

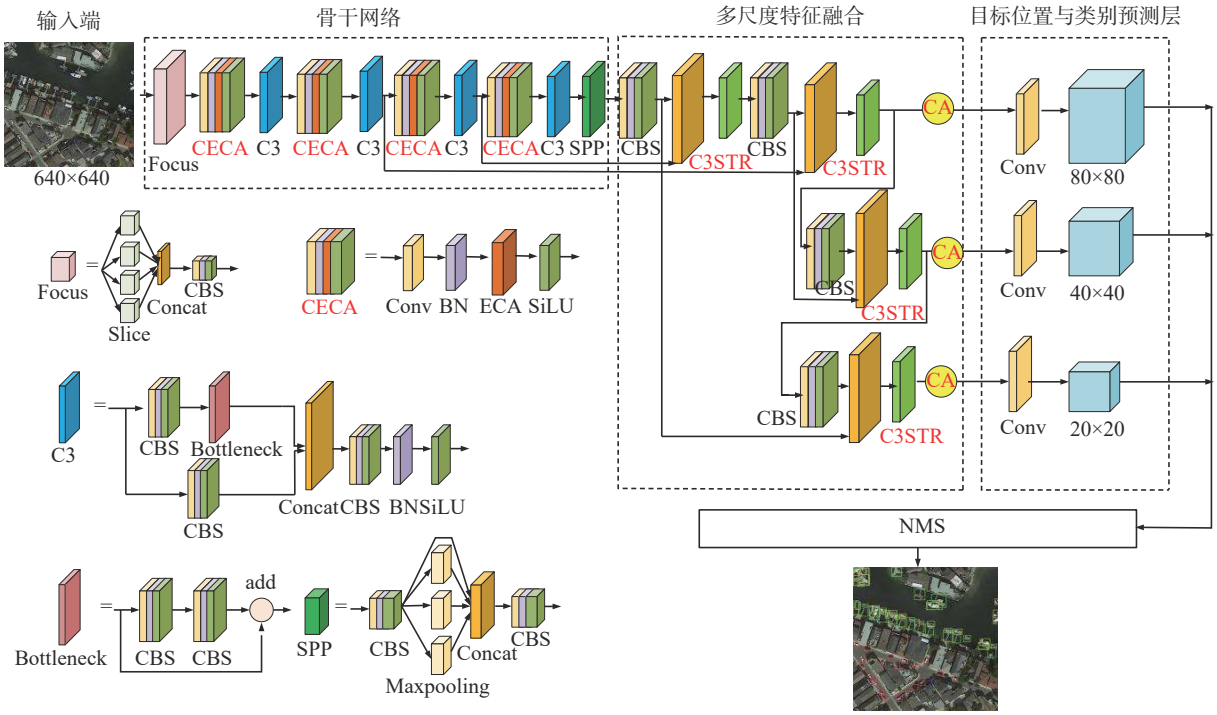


图 1　改进 YOLOv5s 的遥感图像目标检测算法网络结构图

Fig. 1　Network structure of remote sensing image object detection algorithm with improved YOLOv5s

2.1　卷积块通道注意力模型

由于输入的遥感图像背景信息复杂，在进行多次卷积操作时，背景的迭代累积会造成大量的冗余信息，从而掩盖待检测目标，导致目标检测平均精确度降低。因此本文在 YOLOv5s 骨干网络的卷积块中加入 ECA[20]，可以减少全连接层的冗余信息，更加精确的识别目标位置和类别信息，其结构如图 2 所示。

首先，对输入的遥感图像的特征图使用全局平均池化 (global average pooling, GAP) 来获得通道权重，得到 1×1×C 的特征图，进而考虑到每个通道的最重要的知识；然后，使用包含 k 个参数的快速一维卷积来捕获局部跨通道的交互信息，并经过 Sigmoid 层输出新的 1×1×C 的特征图；最后，与输入的特征图元素进行相乘，得到更新后的特征图。ECA[20] 与 SE[21] 和 CBAM[17] 中的通道注意力模块相比，采用局部卷积运算，减少了计算复杂度，性能获得明显提升。

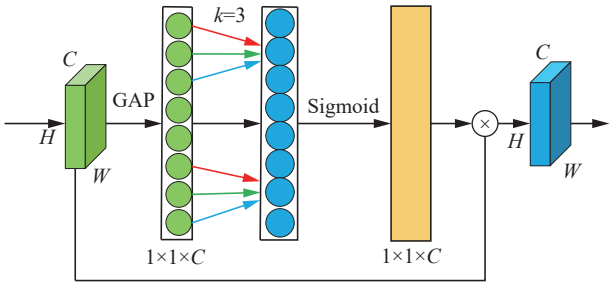


图 2　ECA 模块
Fig. 2　ECA module

2.2　多尺度特征融合

2.2.1　加权双向特征融合

本文在特征金字塔中进行了简化的跨尺度连接和上下文信息加权操作[15]，如图 3 所示。红色虚线框为本文使用的特征金字塔模型，主要分为四部分：自顶向下的过程（绿色箭头）、跨尺度连接（蓝色箭头）、上下文信息加权（对红色区域内部的 N_3、N_4、N_5 的输入流进行加权）和自底向上的过程（橙色箭头）。

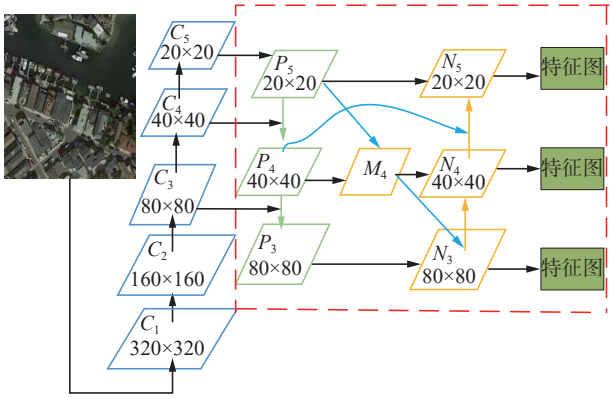


图 3　加权双向特征融合
Fig. 3　Weighted bidirectional feature fusion

具体步骤如下：

当输入图像尺寸为 640×640 时，依次经过多次卷积下采样操作之后提取到不同尺寸的特征图 C_1、C_2、C_3、C_4、C_5。从主干网络获取到最后 3 个特征层 C_3、C_4、C_5 来进行加权双向特征融合网络的构建。自顶向下的过程如下：对特征图 C_5 使用 1×1 卷积调整通道数后得到特征图 P_5；然后对 P_5 进行 2 倍上采样与 C_4 进行特征融合，使用跨阶段局部网络进行特征提取获得特征层 P_4。对 P_4 进行 2 倍上采样后与 C_3 进行特征融合，使用跨阶段局部网络进行特征提取获得特征层 P_3。

选取 M_4 作为特征图 P_4 和 P_5 的映射图，通过 3×3 卷积实现跨尺度连接。可以观察到：N_3 的输入流有 P_3 和 M_4，N_4 的输入流有 P_4、M_4 和 N_3，N_5 的输入流有 P_5 和 N_4。使用快速归一化的融合方法[15] 对 N_3、N_4、N_5 的输入流分配相应的权重信息，来学习不同特征的重要性。同时结合自顶向下和自底向上两个过程的特征图之间的交互信息，输出最终更新后得到的特征图。

快速归一化的融合方法[15] 训练速度快、效率高，其计算公式为

$$O=\sum_i \frac{\omega_i \cdot X_i}{\varepsilon+\sum_j \omega_j}$$

式中：O 表示输出特征图；ω_i 表示该层特征图的权重系数；X_i 表示需要进行融合的特征图；$\varepsilon \leqslant 0.001$。

以图 3 中的特征图 C_4 为例，式（1）和式（2）描述了相邻两层特征图在第 4 层的融合情况：

$$M_4=\mathrm{Conv}\left(\frac{\omega_1 \cdot P_4^{\mathrm{in}}+\omega_2 \cdot \mathrm{Resize}\left(P_5^{\mathrm{in}}\right)}{\omega_1+\omega_2+\varepsilon}\right) \tag{1}$$

$$N_4=\mathrm{Conv}\left(\frac{\omega_1' \cdot P_4^{\mathrm{in}}+\omega_2' \cdot P_4^{\mathrm{td}}+\omega_3' \cdot \mathrm{Resize}\left(P_3^{\mathrm{out}}\right)}{\omega_1+\omega_2+\omega_3+\varepsilon}\right) \tag{2}$$

式中：M_4 是自顶向下融合的第 4 层的中间特征；N_4 是自底向上融合的第 4 层的中间特征；ω_i 和 ω_i' 为不同层级特征图的权重；P_i^{in} 和 P_i^{out} 分别表示第 i 层特征图的输入和自底向上融合过程中第 i 层特征图的输出；P_i^{td} 表示自顶向下融合过程中第 i 层特征图中间节点的输出；Resize 指通过上采样或下采样操作使输入特征图的尺寸保持一致的操作；Conv 通常是一个用于特征处理的卷积操作；ε 为参数，此处设为 0.000 1。

自底向上的过程如下：对特征层 N_3 使用 3×3 卷积进行 2 倍下采样后与 M_4 堆叠，将融合后特征图进行 3×3 卷积操作消除混叠效应，同时使用跨阶段局部网络对 N_4 进行特征提取，此时获得的特征层为 40×40×512。同理，将特征层 N_4 使用 3×3 卷积进行 2 倍下采样后与 P_5 进行融合并消除混叠效应，使用跨阶段局部网络对 N_5 进行特征提取，获得的特征层为 20×20×1 024。

2.2.2　C3STR 结构

随着网络结构的加深，经过多次卷积操作，遥感图像中小目标应该具有的大部分目标特征信息在高级特征图中丢失。所以，在特征融合部分借鉴了 Swin Transformer[18] 的思想，将其嵌入到 C_3 卷积块中，使用 C3STR 结构作为一个辅助模块，引入一些 Transformer 的离散参数，借助窗口自注意力模块增强小目标的语义信息和特征表示，改进后的卷积结构如图 4 所示。

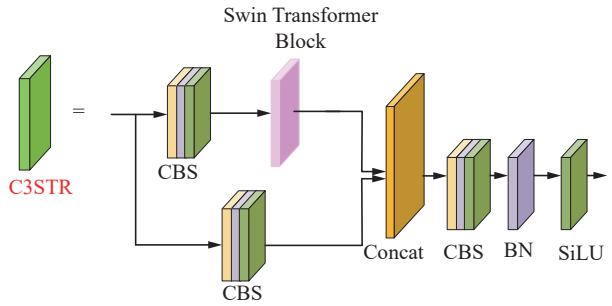


图 4　C3STR 结构

Fig. 4　C3STR structure

图 4 中 Swin Transformer Block (STB)[18] 由成对的窗口多头自注意力模块 (window multi-head self-attention, W-MSA)、滑动窗口多头自注意力模块 (shifted window multi-head self-attention, SW-MSA) 和多层感知机 (multi-layer perceptron, MLP) 构成，并且每个模块内部采用残差连接。其中局部窗口大小为 7，多层感知机隐藏层的嵌入维度为 4。多头自注意力机制的计算过程如下：

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{SoftMax}\left(\boldsymbol{Q}\boldsymbol{K}^{\text{T}}/\sqrt{d}+B\right)\boldsymbol{V}$$

式中：Attention 表示注意力；SoftMax 表示归一化指数函数；$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 分别为查询、键和值矩阵；d 为输入特征图的通道数；B 为相对位置偏差。通过引入 B，能够带来明显的提升效果。

与传统的 Transformer 中的多头自注意力模块相比，C3STR 模块中以划分局部窗口的方式控制每一个窗口中计算区域实现跨窗口的信息交互，降低计算复杂度和网络计算量。

2.2.3　坐标注意力机制

使用坐标注意力机制 (coord attention, CA)[22] 对融合后的特征图进行更新，能够更好地抑制连续多次下采样过程中出现的目标信息丢失严重的情况，得到具有方向感知和位置感知信息的特征图，如图 5 所示。

首先，将双向跨尺度连接和上下文信息加权操作中学习到的特征图在水平和垂直两个方向上进行全局平均池化操作，得到 C 个 H×1 大小的特征图和 C 个 1×W 大小的特征图，并对每个通道在空间信息上进行编码，同时获取通道信息和位置信息。然后，在空间维度上进行拼接，使用 1×1 卷积来对信道进行压缩，得到 C/r×1×(W+H) 大小的特征图；再利用两个 1×1 卷积将水平和垂直两个方向的特征图变换到和输入特征图相同的通道数，并使用 Sigmoid 激活函数，得到每个通道对应的权重信息。最后，通过归一化加权与输入特征图进行相乘得到新的特征映射图。

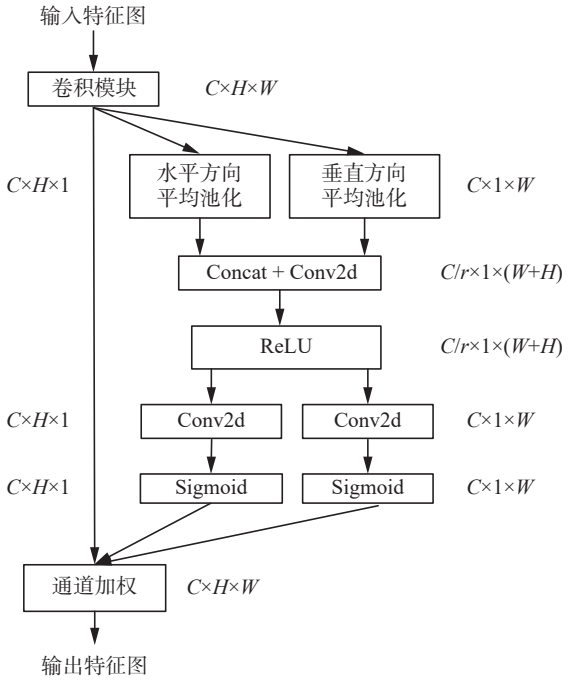


图 5　坐标注意力机制

Fig. 5　CoordAttention mechanism

2.3　损失函数

本文算法的损失函数主要由 3 部分组成，分别是置信度损失、分类损失和定位损失。置信度损失和分类损失均采用二分类交叉熵损失函数 (BCE Loss) 进行计算；置信度损失计算的是所有样本的置信度损失，分类损失只计算正样本的分类损失。回归损失采用 CIoU Loss 来计算，且只计算正样本的回归损失，计算公式为

$$L_{\text{loss}} = \lambda_1 l_{\text{obj}} + \lambda_2 l_{\text{cls}} + \lambda_3 l_{\text{box}}$$

$$l_{\text{obj}} = -\sum_{i=1}^{N} \hat{p}_i \ln(p_i) + \left(1-\hat{p}_i\right)\ln(1-p_i)$$

$$l_{\text{cls}} = -\sum_{i=1}^{N} \hat{y}_i \ln(y_i) + \left(1-\hat{y}_i\right)\ln(1-y_i)$$

$$l_{\text{box}} = 1 - I_{\text{oU}} + \frac{\rho^2}{c^2} + \alpha\upsilon$$

式中：l_{obj}表示置信度损失；l_{cls}表示分类损失；l_{box}表示定位损失；λ_1、λ_2、λ_3为平衡系数，其中，在置信度损失中，对 3 个不同的大中小预测特征层分配不同的权重来平衡不同尺度特征图的损失，其权重系数分别为 0.4、1.0 和 4.0；p_i、y_i分别表示置信度真实值和某一类别的真实概率值，$\hat{p}_i$和$\hat{y}_i$分别表示通过 Sigmoid 函数得到的置信度预测值和某一类别的预测概率值；I_{oU} 表示预测框和真实框的交并比；ρ 为预测框和真实框的中心点距离；c 为预测框和真实框的最小包围矩形的对角线长度；α为υ的影响因子；υ为预测框和真实框的宽高比相似度。

3 实验结果

3.1 实验环境以及参数设置

本实验使用的操作系统为 Ubuntu 18.04 LTS, GPU 为 NVIDIA GeForce RTX 2080Ti, CUDA 为 10.2, 深度学习框架为 pytorch1.10。消融实验中采用随机梯度下降算法 (stochastic gradient descent, SGD) 训练 300 epoch。初始学习率为 0.01, 最小学习率为 0.000 1, batchsize 为 32。在模型的初始训练中, 首先进行 3 个 epoch 的 warm-up 训练, 此时 SGD 的动量参数设置为 0.8。之后根据如下语句自适应调整学习率: Init_lr =max(batch_size / 64* Init_lr, 1e-4), Min_lr = max(batch_size / 64 * Min_lr, 1e-6)。其中 Init_lr 代表初始学习率, Min_lr 代表最小学习率, batchsize 为当前设置训练过程中设置的大小。 动量参数和权重衰减分别为 0.937 和 0.000 5。

3.2 数据集和评价指标

本文所使用的数据集为 DOTA 数据集和 RSOD 数据集。DOTA 数据集有 2 806 张图片, 188 282 个实例, 共 15 个类别。我们采用了图像切割的方式对数据集进行预处理, 将原始图像按照 gap=200, subsize=1 024 的方式裁剪成多个分辨率为 1 024×1 024 的子图像, 对于自身分辨率不足 1 024×1 024 的图像, 通过填充的方式将其填充为 1 024×1 024 大小。裁剪之后的图片共 21 046 张, 从中随机选取 15 749 张图片作为训练集, 5 297 张图片作为测试集, 在本文中我们仅对其进行水平框的任务检测。本实验中所使用的 RSOD 数据集共 936 张图片, 包含 4 个类别, 即飞机、油桶、立交桥、操场, 从中随机选取 742 张图片作为训练集, 剩余的 194 张图片作为测试集。

本文采用平均精度 (average precision, AP)、平均精度均值 (mean average precision, mAP)、检测速率作为评价指标。

3.3 实验结果及分析

3.3.1 DOTA 数据集实验结果及分析

通过改进 YOLOv5s 模型, 本文在 DOTA 数据集上进行了消融实验, 来证明改进之后模型的有效性, 实验结果如表 1 所示。其中 CECA 表示在卷积块中加入 ECA 之后形成的新的卷积块; CA 表示坐标注意力机制; WFPN 表示加权双向特征融合; C3STR 表示引入具有 Swin Transformer 结构特性的网络模块。

表 1　本文算法在 DOTA 数据集的消融实验结果比较

Table 1　Comparison of ablation experimental results of the algorithm in this paper on the DOTA dataset　%

方法	mAP
YOLOv5s	69.5
YOLOv5s+CECA	71.4
YOLOv5s+CA	71.2
YOLOv5s+CECA+CA	72.0
YOLOv5s+WFPN	71.0
YOLOv5s+C3STR	71.6
YOLOv5s+WFPN+C3STR	71.8
YOLOv5s+CA+WFPN+C3STR	72.8
YOLOv5s+CECA+CA+WFPN+C3STR	73.6

由表 1 可知, 本文提出的改进方法中, 在 YOLOv5s 骨干网络中加入 CECA, mAP 可以提升 1.9%; 加入 CA, mAP 可以提升 1.7%; 加入 WFPN, mAP 提升 1.5%; 加入 C3STR, mAP 可以提升 2.1%。当所有改进的方法同时加入原始 YOLOv5s 模型后, 整体的 mAP 值提升了 4.1%, 此时 mAP 为 73.6%。

将本文改进的模型与其他主流目标检测算法在 DOTA 数据集上进行了实验对比, 实验结果如表 2 所示。

由表 2 可知, 本文改进的模型 Swin-YOLOv5s 相比两阶段目标检测算法 Faster R-CNN, mAP 和检测速率有大幅度提升。与单阶段目标检测算法 SSD、RetinaNet、YOLOv3、YOLOv4、YOLOv5s、FMSSD 相比, mAP 值分别提升了 21.2%、12.0%、9.1%、5.2%、4.1% 和 1.2%。此外, 可以看出本文模型具有较高的检测速度。

基于 Transformer 改进的 YOLOv5 模型 TPH-YOLOv5 在 VisDrone2021 数据集上有效提高了小目标的平均检测精度值, 为了验证其有效性, 本文在 DOTA 数据集上进行实验, 可以发现各个类别的 AP 值均有不同程度的提升, 但是速度明显变慢。本文模型与 TPH-YOLOv5 相比, mAP 提升了 2.2%。ViT-YOLO 基于 YOLOv4-P7 模型, 在骨干网络的末端引入带有多头注意力机制的卷积, 同时使用具有 5 个检测头的 BiFPN 网络进行多尺度特征融合, 经过在 DOTA 数据集上进行实验验证, mAP 达到了 73.1%。而在本文中使用的加权双向特征融合是经过简化处理之后的 BiFPN 网络, 具有 3 个检测头部, 可以在不引入过多参数量的情况下实现精度和速度的提升。本文模型与 ViT-YOLO 相比, mAP 提升了 0.5%。

表 2　不同算法在 DOTA 数据集的检测结果比较

Table 2　Test results of different algorithms on the DOTA dataset

算法	各类别准确率/%															mAP/%	检测速率/(f·s^{-1})
	飞机	棒球场	桥梁	田径场	小型车辆	大型车辆	船舰	网球场	篮球场	储油罐	足球场	环岛	港口	游泳池	直升机		
SSD[23]	80.9	70.3	18.2	68.7	22.0	58.4	34.6	88.0	61.2	23.5	65.3	32.5	70.8	38.4	53.5	52.4	38
Faster R-CNN[3]	74.7	66.4	14.0	63.7	8.8	38.0	13.2	84.6	53.2	17.4	57.3	28.2	56.3	25.7	27.8	42.0	8
RetinaNet[24]	86.2	79.6	26.0	68.9	29.7	62.0	50.7	94.3	67.5	29.2	66.9	55.2	70.8	69.7	66.7	61.6	15
YOLOv3[6]	90.5	53.0	27.2	47.7	68.3	69.9	82.7	94.3	52.8	60.7	50.7	31.6	77.8	78.6	82.0	64.5	18
YOLOv4[7]	94.2	77.1	42.1	42.8	70.8	71.1	88.3	94.6	55.9	68.5	42.9	38.7	81.9	74.0	83.8	68.4	17
YOLOv5s	91.5	73.4	43.9	63.6	63.0	84.9	87.0	93.0	63.6	66.9	51.3	59.2	83.0	61.0	56.9	69.5	38
FMSSD[25]	89.1	81.5	48.2	67.9	69.2	73.6	76.9	90.7	82.7	73.3	52.7	67.5	72.4	80.6	60.2	72.4	16
TPH-YOLOv5[16]	91.8	77.7	49.5	68.1	66.6	84.7	87.2	93.7	64.7	69.7	53.1	62.7	84.1	63.3	53.9	71.4	18
ViT-YOLO[14]	94.7	79.2	48.8	60.7	68.4	72.7	89.1	94.8	58.8	70.2	53.1	57.9	84.0	77.8	85.8	73.1	16
Swin-YOLOv5s	93.4	79.5	50.6	67.2	69.6	89.2	88.6	94.5	67.4	71.2	56.2	62.9	85.7	65.6	62.9	73.6	23

由表 2 可知，本文算法相比于原始的 YOLOv5s，在小目标较多的类别中，如：桥梁、小型车辆、大型车辆、船舰和储油罐等，在 DOTA 数据集上的 mAP 分别提升了 6.7%、6.6%、4.3%、1.6% 和 4.3%。可以看出，本文算法对小目标检测的改进效果有一定程度的提升。与通用的目标检测算法 SSD、Faster R-CNN、RetinaNet 和 YOLOv3 相比，桥梁、小型车辆、大型车辆、船舰和储油罐等小目标的精度值均有明显的提升。

此外，本文算法还和最新改进的算法进行了比较，可以发现：与 TPH-YOLOv5 相比，桥梁、小型车辆、大型车辆、船舰和储油罐等小目标的精度值分别提升了 1.1%、3.0%、4.5%、1.4% 和 1.5%，均有一定程度的提升。与 YOLOv4、FMSSD 和 ViT-YOLO（基于 YOLOv4-P7 模型）等这些网络模型本身较为复杂的算法相比，本文算法以轻量级的 YOLOv5s 作为基础模型进行改进，在桥梁、小型车辆、大型车辆、船舰和储油罐等小目标检测结果的精度值仍然具有较好的表现，且检测速度也优于这些算法。

3.3.2　RSOD 数据集实验结果及分析

为使本文算法具有更好的泛化性，将本文算法在 RSOD 数据集上进行消融实验，如表 3 所示。由表 3 可知，本文算法在 RSOD 数据集上也具有一定的泛化性，mAP 提升了 5.3%。

表 3　本文算法在 RSOD 数据集的消融实验结果比较

Table 3　Ablation studies on the RSOD dataset　%

方法	mAP
YOLOv5s	83.6
YOLOv5s+CECA	84.6
YOLOv5s+CA	85.1
YOLOv5s+CECA+CA	85.6
YOLOv5s+WFPN	84.5
YOLOv5s+C3STR	84.8
YOLOv5s+WFPN+C3STR	86.5
YOLOv5s+CA+WFPN+C3STR	87.6
YOLOv5s+CECA+CA+WFPN+C3STR	88.9

同时将本文算法与其他目标检测算法进行实验对比，结果表明该算法仍然具有较高的平均准确率和速率，实验结果如表 4 所示。由表 4 可知，与原始的 YOLOv5s 相比，各个类别的精度值均有所提升。虽然，本文算法和其他算法相比，并不能保证各个类别的精度值均是最高的，但是这并不影响总体的平均精确度高于其他算法，本文算法可以在保证较高精确度的同时达到较高的速度，这也算是一种有意义的改进。

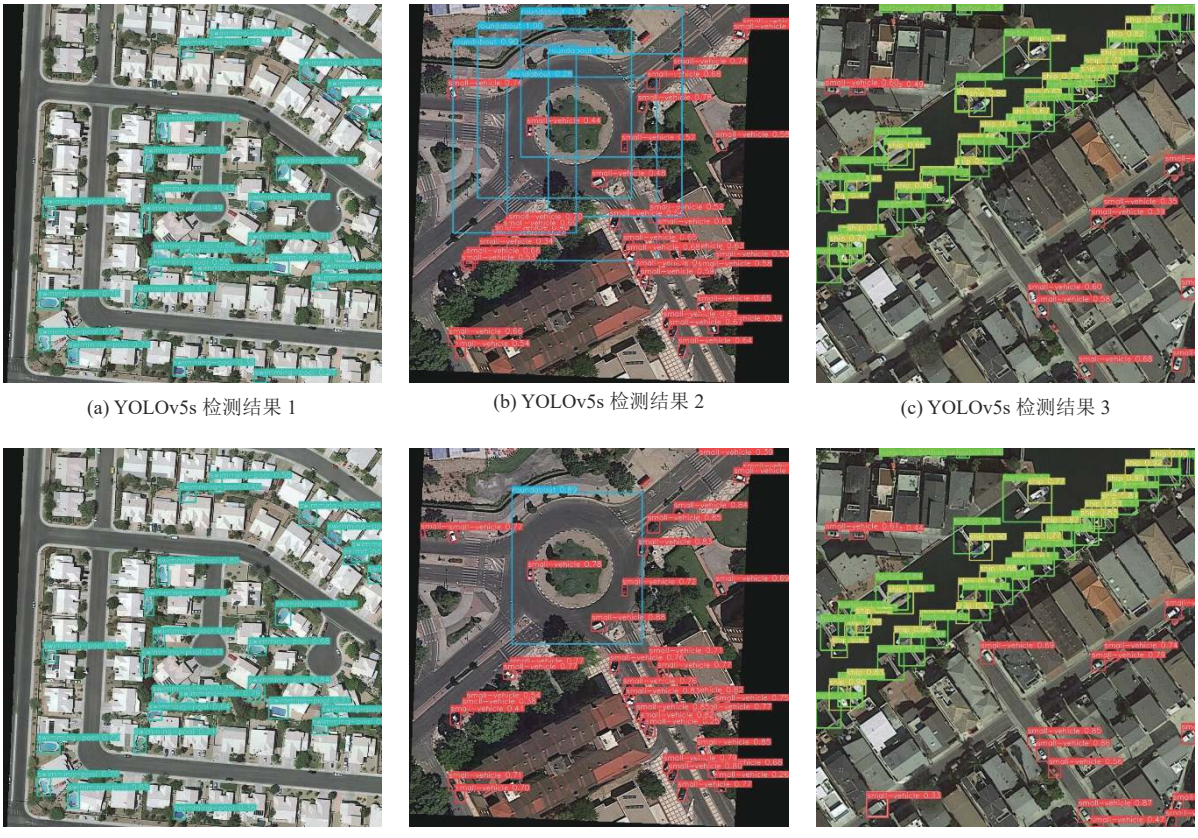
本文对 DOTA 数据集上的测试结果进行了可视化展示，如图 6 所示。从可视化结果中可以发

现, 本文算法的检测框与待检测的遥感目标贴合的更为紧密, 同时能够察觉出一些不易被发现的遥感小目标, 在一定程度上降低了小目标的漏检率, 进而提升了遥感目标的检测精度值。

表 4　不同算法在 RSOD 数据集的检测结果比较

Table 4　Comparison of detection results of different algorithms on the RSOD dataset

算法	骨干网络	各类别准确率/%				mAP/%	检测速率/(f·s^{-1})
		飞机	油桶	立交桥	操场		
SSD[23]	VGG16	52.1	96.6	56.7	100.0	76.4	46
Faster R-CNN[3]	VGG16	63.1	84.1	76.9	97.8	80.5	7
YOLOv3[6]	Darknet53	62.2	95.1	70.4	98.6	81.6	30
YOLOv4[7]	CSPDarknet53	81.3	98.1	71.7	100.0	87.8	28
YOLOv5s	CSPDarknet	89.7	79.4	71.4	94.0	83.6	48
Swin-YOLOv5s	Modified CSPDarknet	90.4	85.8	81.5	97.9	88.9	35



(a) YOLOv5s 检测结果 1　(b) YOLOv5s 检测结果 2　(c) YOLOv5s 检测结果 3

(d) Swin-YOLOv5s 检测结果 1　(e) Swin-YOLOv5s 检测结果 2　(f) Swin-YOLOv5s 检测结果 3

图 6　在 DOTA 数据集上的可视化检测结果

Fig. 6　Visualization detection results on the DOTA dataset

4　结束语

针对遥感图像中目标检测存在的问题, 本文提出了 Swin-YOLOv5s 算法。首先, 在骨干特征提取网络的卷积块中加入 ECA, 通过避免降低通道维度来学习有效的通道注意力。其次, 在自顶向下和自底向上的多尺度特征融合过程中, 进行有效的交叉连接和上下文信息加权操作, 提高待检测目标的信息提取能力。最后, 在特征融合的过程中引入 C3STR 模块和坐标注意力机制, 增强小目标的全局感知能力, 对小目标产生更好的拟合效果。经过实验对比, 本文算法相比于原始的 YOLOv5s, 平均检测准确率在 DOTA 数据集和 RSOD 数据集上分别提升了 4.1% 和 5.3%, 由此表明本文算法在遥感图像目标检测领域的有效性。

但是，本文算法也存在的一定的局限性：该算法会导致网络结构变复杂，在提升遥感目标检测精度的同时会使网络的推理时间增加，检测速度下降，实时性变差。未来，我们会继续探索轻量级骨干特征提取网络和新的特征融合方式来简化网络结构，实现遥感目标检测高速度和高精度的平衡，进而实现实时检测在工业场景中的应用。

参考文献：

[1] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014: 580−587.

[2] GIRSHICK R. Fast R-CNN[C]//2015 IEEE International Conference on Computer Vision. Santiago: IEEE, 2015: 1440−1448.

[3] REN Shaoqing, HE Kaiming, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 39(6): 1137–1149.

[4] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 779−788.

[5] REDMON J, FARHADI A. YOLO9000: better, faster, stronger[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 6517−6525.

[6] FARHADI A, REDMON J. YOLOv3: an incremental improvement[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 1804−2767.

[7] BOCHKOVSKIY A, WANG C Y, LIAO H M, et al. YOLOv4: optimal speed and accuracy of object detection[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 2−7.

[8] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//2017 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 936−944.

[9] LIU Shu, QI Lu, QIN Haifang, et al. Path aggregation network for instance segmentation[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 8759−8768.

[10] 林娜, 冯丽蓉, 张小青. 基于优化 Faster-RCNN 的遥感影像飞机检测 [J]. 遥感技术与应用, 2021, 36(2): 275–284.
LIN Na, FENG Lirong, ZHANG Xiaoqing. Aircraft detection in remote sensing image based on optimized faster-RCNN[J]. Remote sensing technology and application, 2021, 36(2): 275–284.

[11] 姚艳清, 程塨, 谢星星, 等. 多分辨率特征融合的光学遥感图像目标检测 [J]. 遥感学报, 2021, 25(5): 1124–1137.
YAO Yanqing, CHENG Gong, XIE Xingxing, et al. Optical remote sensing image object detection based on multi-resolution feature fusion[J]. National remote sensing bulletin, 2021, 25(5): 1124–1137.

[12] 张晓雅, 李承政, 徐静杉, 等. 级联结构的遥感目标检测算法 [J]. 计算机辅助设计与图形学学报, 2021, 33(10): 1524–1531.
ZHANG Xiaoya, LI Chengzheng, XU Jingshan, et al. Cascaded object detection algorithm in remote sensing imagery[J]. Journal of computer-aided design & computer graphics, 2021, 33(10): 1524–1531.

[13] 李婕, 周顺, 朱鑫潮, 等. 结合多通道注意力的遥感图像飞机目标检测 [J]. 计算机工程与应用, 2022, 58(1): 209–217.
LI Jie, ZHOU Shun, ZHU Xinchao, et al. Remote sensing image aircraft target detection combined with multiple channel attention[J]. Computer engineering and applications, 2022, 58(1): 209–217.

[14] ZHANG Zixiao, LU Xiaoqiang, CAO Guojin, et al. ViT-YOLO: transformer-based YOLO for object detection[C]//2021 IEEE/CVF International Conference on Computer Vision Workshops. Montreal: IEEE, 2021: 2799−2808.

[15] TAN Mingxing, PANG Ruoming, LE Q V. EfficientDet: scalable and efficient object detection[EB/OL]. (2019−11−20) [2022−02−12].https://arxiv.org/abs/1911.09070.

[16] ZHU Xingkui, LYU Shuchang, WANG Xu, et al. TPH-YOLOv5: improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios[C]//2021 IEEE/CVF International Conference on Computer Vision Workshops. Montreal: IEEE, 2021: 2778−2788.

[17] WOO S, PARK J, LEE J Y, et al. CBAM: convolutional block attention module[M]//Computer Vision - ECCV 2018. Cham: Springer International Publishing, 2018: 3−19.

[18] XU Xiangkai, FENG Zhejun, CAO Changqing, et al. An improved swin transformer-based model for remote sensing object detection and instance segmentation[J]. Remote sensing, 2021, 13(23): 4779.

[19] LIU Ze, LIN Yutong, CAO Yue, et al. Swin transformer: hierarchical vision transformer using shifted windows[C]//2021 IEEE/CVF International Conference on Computer Vision. Montrea: IEEE, 2021: 9992−10002.

[20] WANG Qilong, WU Banggu, ZHU Pengfei, et al. ECA-net: efficient channel attention for deep convolutional neural networks[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 11531−11539.

[21] HU Jie, SHEN Li, SUN Gang. Squeeze-and-excitation networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 7132−7141.

[22] HOU Qibin, ZHOU Daquan, FENG Jiashi. Coordinate attention for efficient mobile network design[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 13708−13717.

[23] LIU Wei, ANGUELOV D, ERHAN D, et al. SSD: Single Shot MultiBox Detector[C]//European Conference on Computer Vision. Cham: Springer, 2016: 21−37.

[24] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]//2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 2999−3007.

[25] WANG Peijin, SUN Xian, DIAO Wenhui, et al. FMSSD: feature-merged single-shot detection for multiscale objects in large-scale remote sensing imagery[J]. IEEE transactions on geoscience and remote sensing, 2020, 58(5): 3377–3390.

作者简介:

赵文清, 教授, 博士, 主要研究方向为人工智能与图像处理。获河北省科技进步二等奖、三等奖各 1 项。发表学术论文 50 余篇。

康怿瑾, 硕士研究生, 主要研究方向为深度学习与目标检测。

赵振兵, 教授, 博士, 主要研究方向为电力视觉。主持国家自然科学基金等纵向课题 10 项。获省科技进步一等奖 1 项 (第三完成人)。以第一完成人获得国家专利授权 16 项; 以第一作者出版专著 2 部、发表学术论文 50 余篇。