DOI:10.19651/j.cnki.emt.2212528

基于加权感受野和跨层融合的遥感小目标检测*

张绍文[1] 史卫亚[2] 张世强[1] 王甜甜[1]
(1.河南工业大学信息科学与工程学院 郑州 450001;2.河南工业大学人工智能与大数据学院 郑州 450001)

摘 要: 针对遥感图像中小目标特征易丢失、易受背景噪声影响和定位难的问题,本文对 YOLOX-S 目标检测模型进行改进。使用二维离散余弦变换对 CBAM 注意力模块进行改进并加入到主干网络当中,提高网络对小目标的关注度;其次提出一种加权多重感受野空间金字塔池化模块,提升模型对多尺度目标尤其是小尺度目标的感知能力;采用跨层特征融合的思想,提出一种跨层注意力融合模块,使深层结构中尽可能保留更多的小目标特征;最后使用 EIoU 损失加强对小目标的定位能力。由大量实验分析可知,在 RSOD 数据集上,改进模型的 APs 值相对于基准模型提高了 5.1%,在 DIOR 数据集上提高了 2.4%,参数量仅增加了 1.01 M,检测速度达到 93.6 fps,满足实时性的检测要求。此外,相对于当前最新的目标检测模型,本文改进模型也具有一定优势。

关键词: 图像处理;遥感小目标;YOLOX;跨层融合;离散余弦变换
中图分类号: TP391.4 **文献标识码:** A **国家标准学科分类代码:** 510.4050

Remote sensing small target detection based on weighted receptive field and cross-layer fusion

Zhang Shaowen[1] Shi Weiya[2] Zhang Shiqiang[1] Wang Tiantian[1]
(1. College of Information Science and Engineering, Henan University of Technology, Zhengzhou 450001, China;
2. College of Artificial Intelligence and Big Data, Henan University of Technology, Zhengzhou 450001, China)

Abstract: Aiming at the problems that small target features in remote sensing images are easily lost, easily affected by background noise and difficult to locate, this paper improves the YOLOX-S target detection model. Firstly, the CBAM is improved by using the two-dimensional discrete cosine transform and added to the backbone network to improve the attention of the network to small targets; secondly, a weighted multi-receptive spatial pyramid pooling module is proposed to improve the perception ability of the model to multi-scale targets, especially to small-scale targets. Thirdly, using the idea of cross-layer feature fusion, a cross-layer attention fusion module is proposed to retain as many small target features as possible in the deep structure; finally, EIoU loss is used to enhance the localization ability of small targets. As shown by extensive experimental analysis, the APs value of the improved model improves by 5.1% relative to the baseline model on the RSOD dataset and by 2.4% on the DIOR dataset, and the number of parameters increases by only 1.01 M. The detection speed reaches 93.6 fps, which meets the detection requirements of real-time. In addition, the improved model in this paper also has certain advantages over the current state-of-the-art target detection models.

Keywords: image processing;remote sensing small target;YOLOX;cross-layer fusion;discrete cosine transform

0 引 言

随着航天遥感技术的发展,遥感图像的分类、分割、检测和跟踪等任务成为了当前图像处理领域的热点。近年,基于深度学习的目标检测算法在诸多领域取得了显著的成就[1]。尽管这些目标检测算法对于中大型目标有着不错的检测效果,但在遥感目标检测任务中对于尺寸小、像素值少的小目标(基于相对尺度的定义,将目标边界框的宽高与图像的宽高比例小于 0.1 的目标定义为小目标)仍存在着挑战。

针对通用目标检测算法对于遥感小目标检测困难的问题,研究人员先后提出了大量的遥感小目标检测方法。牛浩青等[2]在 YOLOv3 的基础上,引入了一种基于门控通道

收稿日期:2022-12-31
* 基金项目:国家自然科学基金(62006071)、河南省科技攻关项目(212102210149)资助

注意力机制和自适应上采样的方法，提升了模型对于遥感中小型目标的检测能力。汪鹏等[3]在模型中构造调制的特征自适应网络，提取更丰富的目标特征，同时引入上下文特征金字塔模块用来解决高层语义信息与感受野之间的矛盾，进一步提高了检测的精度。唐建宇等[4]通过优化 YOLOv5 的主干部分和加入注意力机制，优化模型对遥感小目标的特征提取能力。尽管以上方法在针对遥感小目标的检测中取得了不错的效果，但仍然有可提升的空间。目前对于遥感小目标检测的难点在于以下几个方面：1）目标特征信息过少且易丢失；2）小目标特征易受背景噪声影响景复杂；3）相对于中大型目标小目标更难以定位。

对于以上遥感图像中小目标检测的难点，本文以 YOLOX-S 目标检测模型作为基线模型，提出一种针对于遥感小目标的改进模型。其主要方法为：1）在主干网络中加入改进的卷积注意力机制模块（convolutional block attention module，CBAM）[5]，过滤掉由复杂背景引起的噪声；2）提出加权多重感受野空间金字塔池化模块，加强对小目标的检测能力；3）采用多尺度融合的方案，构建出一种跨层注意力融合模块，将浅层特征和深层特征相融合，使得融合后的特征图中最大程度的保留小目标的信息；4）为了使模型对边界框的回归更准确，采用损失函数（efficient intersection over union，EIoU）[6]。

1　YOLOX-S 目标检测算法

YOLOX[7]是在 2021 年由旷世提出的 YOLO 系列单阶段目标检测模型。YOLOX 有着 X、L、M、S、Tiny 和 Nano 共 6 种型号。相对于其他型号 YOLOX-S 有着较高的精度同时参数量较少，其网络结构如图 1 所示。

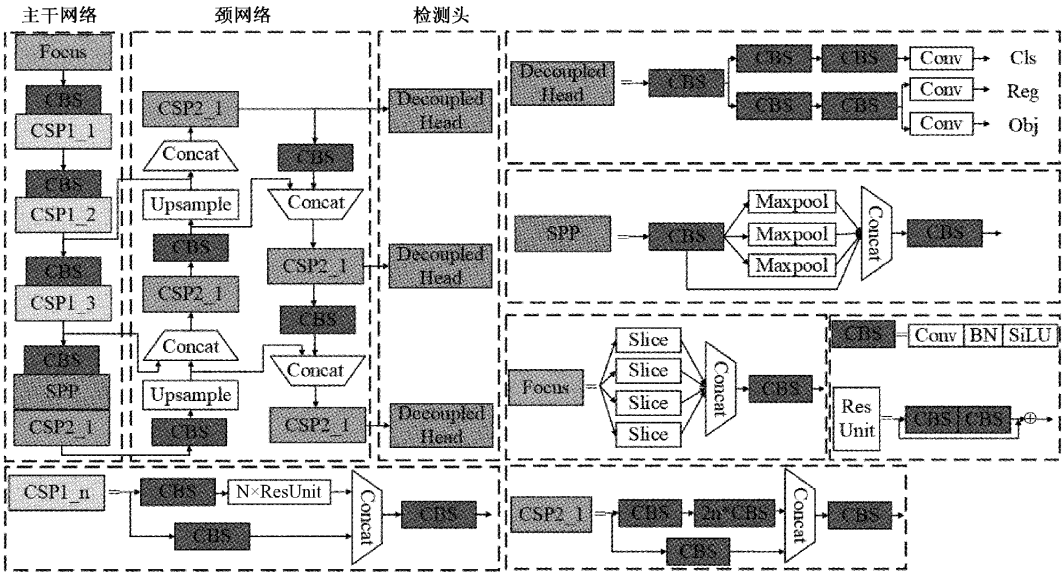


图 1　YOLOX-S 网络结构

YOLOX-S 的主干网络（backbone）采用的是 CSPDarknet，在 Darknet 的基础上引入了跨阶段局部（cross stage partial，CSP）网络结构[8]。CSP（cross stage partial）结构通过切分特征图和残差连接，使得模型在提高精度的同时减少了参数量。YOLOX-S 的颈网络（neck）部分采用的是特征金字塔网络（feature pyramid networks，FPN）[9]和路径聚合网络（path aggregation network，PAN）[10]相结合的结构，通过自顶向下和自底向上的融合连接，最终输出 3 个金字塔特征 $\boldsymbol{X}_i \in \mathbb{R}^{H_i \times W_i \times C}$，其中 $i \in \{1,2,3\}$。YOLOX-S 的检测头（head）部分采用了解耦的设计方式，并重新采取了 anchor free 的边界框生成方式。

2　本文方法

2.1　改进模型的结构

本文模型结构以 YOLOX-S 作为基线模型，改进后的模型结构如图 2 所示，所提出的改进模块分别是：注意力模块 CBAM-MSCA、加权多重感受野空间金字塔池化模块 WMSimSPPF 和跨层注意力融合模块 CAFM。此外使用 EIoU 损失加强模型对小目标的定位能力。

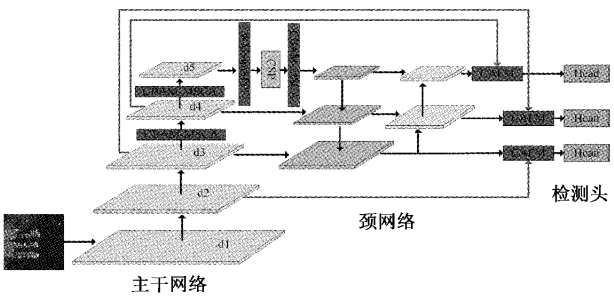


图 2　改进模型的结构

2.2　注意力机制模块 CBAM-MSCA

为了使模型可以更好的过滤背景噪声，加强对小目标特征的关注度，本文在 CBAM 注意力模块的基础上，根据

其存在的问题提出了一个改进的注意力机制模块 CBAM-MSCA。

CBAM 由一个通道注意力模块(channel attention module, CAM)和一个空间注意力模块(special attention module, SAM)级联组成。其中 CAM 分别使用最大值池化和平均值池化来对输入的特征 $\boldsymbol{F} \in \mathbb{R}^{C\times H\times W}$ 进行压缩,目的是获得输入特征所映射的通道信息。将生成的 2 个通道信息特征 $\boldsymbol{F}^c_{\max}, \boldsymbol{F}^c_{\text{avg}} \in \mathbb{R}^{C\times 1\times 1}$ 送入一个由 3 层全连接结构成的多层感知机,最后将得到的特征相加并经 sigmoid 函数处理后得到通道注意力权重 $\boldsymbol{M}_c \in \mathbb{R}^{C\times 1\times 1}$,计算过程如式(1)所示。

$$\boldsymbol{M}_c(\boldsymbol{F}) = \sigma(MLP(AvgPool(\boldsymbol{F})) + MLP(MaxPool(\boldsymbol{F}))) = \sigma(\boldsymbol{W}_1(\boldsymbol{W}_0(\boldsymbol{F}^c_{\text{avg}})) + \boldsymbol{W}_1(\boldsymbol{W}_0(\boldsymbol{F}^c_{\max}))) \tag{1}$$

式(1)中 σ 代表 sigmoid 激活函数,$\boldsymbol{W}_0$ 和 $\boldsymbol{W}_1$ 代表 3 层全连接网络中的权重。

将通道注意力权重 $\boldsymbol{M}_c$ 与输入特征 $\boldsymbol{F}$ 进行内积运算得到融入通道注意力的特征映射 $\boldsymbol{F}_1 \in \mathbb{R}^{C\times H\times W}$。SAM 将 $\boldsymbol{F}_1$ 作为输入,其在通道维度上进行最大值池化和平均值池化,获得 2 个空间信息映射 $\boldsymbol{F}^s_{\max}, \boldsymbol{F}^s_{\text{avg}} \in \mathbb{R}^{1\times H\times W}$,然后将两者进行拼接,经过一个 7×7 的卷积运算和 sigmoid 函数处理后得到空间注意力权重 $\boldsymbol{M}_s \in \mathbb{R}^{1\times H\times W}$,计算过程如式(2)所示。

$$\boldsymbol{M}_s(\boldsymbol{F}_1) = \sigma(f^{7\times 7}([AvgPool(\boldsymbol{F}); MaxPool(\boldsymbol{F})])) = \sigma(f^{7\times 7}([\boldsymbol{F}^s_{\text{avg}}; \boldsymbol{F}^s_{\max}])) \tag{2}$$

式(2)中 σ 代表 sigmoid 激活函数,$f^{7\times 7}$ 为 7×7 的卷积运算。将 $\boldsymbol{M}_s$ 与 $\boldsymbol{F}_1$ 进行内积运算后便得经 CBAM 处理后的特征图。

尽管 CBAM 是一个十分高效的注意力模块但仍然存在着一些缺陷。最主要的是 CAM 中使用全局平均池化进行特征的压缩会造成一定程度特征信息的缺失。文献[11]提出了多频谱通道注意力模块(multi-spectral channel attention module, MSCAM),其从频率的角度证明了 CAM 在计算通道注意力时使用的全局平均值池化本质是离散余弦变换(discrete cosine transform, DCT)的零频分量,然而这显然会忽略掉其他有用的频率分量,尽管在 CAM 中还使用了最大值池化作为频率信息的补充,但这仍然不是最优解。本文基于文献[11],使用二维 DCT 代替 CAM 中的全局平均池化,对 CBAM 进行改进,将频率信息引入到注意力当中,在压缩特征的过程中引入更多有用的频率信息,避免因使用全局平均池化而造成的特征信息的缺失,使网络在特征提取的过程中更好的过滤噪声。将改进后的 CAM 称为 CAM-MSCA,将改进后的 CBAM 称为 CBAM-MSCA,其结构如图 3 所示。

使用二维 DCT 代替全局平均池化的具体做法为:将特征图按通道数均分为多个部分,使用不同频率分量计算各个部分的二维 DCT,其中包括零频分量,即全局平均池化,以压缩更多的频率信息。如图 3 所示,$\boldsymbol{F}^0, \boldsymbol{F}^1, \cdots, \boldsymbol{F}^{n-1}$ 表示输入特征 $\boldsymbol{F} \in \mathbb{R}^{C\times H\times W}$ 被分成的各个部分,其中 $\boldsymbol{F}^i \in \mathbb{R}^{C'\times H\times W}, i \in \{0,1,\cdots,n-1\}, C' = \frac{C}{n}$,$C$ 应当被 n 整除。对于每个部分,分配对应的二维 DCT 频率分量,进行二维 DCT 变换,最后再将各个压缩分量进行拼接得经二维 DCT 压缩的特征图 $\boldsymbol{F}^{2DDCT} \in \mathbb{R}^{C\times 1\times 1}$。计算过程如式(3)~(5)所示。此外,在本文中使用文献[12]中基于启发式两步准则的性能最好的 16 个频率分量。

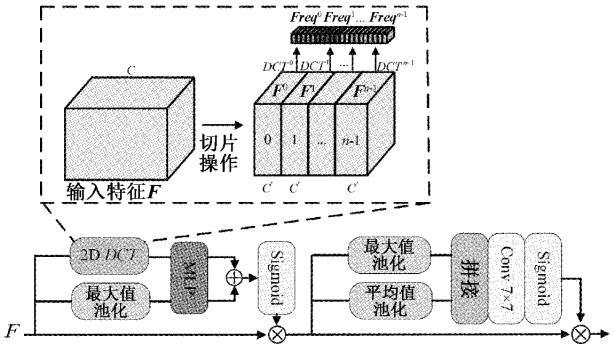


图 3 CBAM-MSCA 的结构

$$\boldsymbol{Freq}^i = 2DDCT^{u_i,v_i}(\boldsymbol{F}^i) = \sum_{h=0}^{H-1}\sum_{w=0}^{W-1} \boldsymbol{F}^i_{:,h,w} B^{u_i,v_i}_{h,w}$$
$$s.t.\ i \in \{0,1,\cdots,n-1\} \tag{3}$$

$$B^{i,j}_{h,w} = \cos\left(\frac{\pi h}{H}\left(i+\frac{1}{2}\right)\right)\cos\left(\frac{\pi w}{W}\left(j+\frac{1}{2}\right)\right) \tag{4}$$

$$\boldsymbol{F}^{2DDCT} = [\boldsymbol{Freq}^0; \boldsymbol{Freq}^1; \cdots; \boldsymbol{Freq}^{n-1}] \tag{5}$$

式(4)为二维 DCT 的基函数,式(3)中 u_i 和 v_i 是预设的对应 $\boldsymbol{F}^i$ 的二维频率分量权重,$\boldsymbol{Freq}^i \in \mathbb{R}^{C'\times H\times W}$ 是各个通道部分经二维 DCT 压缩后的结果。

2.3 加权多重感受野空间金字塔池化模块

在 YOLOX-S 的 backbone 当中使用了空间金字塔池化(spatial pyramid pooling, SPP)[12],其目的是在不同尺度的感受野上提取空间特征信息,提升模型对于空间布局和物体变性的鲁棒性。YOLOv6[13]提出了多感受野特征融合(simplified spatial pyramid pooling-fast, SimSPPF)。其计算量与 SPP 相同,但运算速度快于 SPP。

SimSPPF 级联 3 个大小为 5×5 的最大值池化,得到 3 个感受野分别为 5×5、9×9 和 13×13 的特征图。本文在 SimSPPF 的基础上提出加权多重感受野空间金字塔池化模块(weighted multi-receptive field SimSPPF, WMSimSPPF)。其结构如图 4(a)所示,WMSimSPPF 相对于 SimSPPF 额外构建了 1×1 和 3×3 的小感受野特征图。更小的感受野确保了网络对于小目标特征的捕捉能力,同时多尺度感受野可以更好的使网络获取多尺度物体信息。WMSimSPPF 通过大小为 1×1 的卷积和 2×2 的空洞卷积级联构成额外的 1×1 和 3×3 的小感受野。其中使用空洞卷积构建 3×3 的感受野可以在获取感受野的同时尽可能的保持特征图的分辨率,保留图像边界的细节信息,加强对小尺度目标的检测能力。此外,为了使

WMSimSPPF 能够自适应的学习不同感受野特征图的重要程度，采用加权融合的思想，将不同权重与不同感受野特征图进行乘积后再进行拼接操作。此外，关于权重的生成方式，不再采取传统使用全局平均池化压缩特征图的方式生成，而是采用上节所提到的二维 DCT。具体做法如图 4(b)所示，使用 WMSimSPPF 的输入特征图作为二维 DCT 的输入，特征图经二维 DCT 进行压缩过后，再经 2 层全连接进行特征的提取，最后使用 Sigmoid 函数进行激活得到 6 个特征图权重 $\boldsymbol{W} \in \mathbb{R}^{6\times1\times1}$。将这 6 个权重分别与不同感受野的特征图相乘后再进行下一步操作。以这种方式，进一步加强模型的多尺度感知能力，若模型某个时刻更应该注意小目标，则会赋予小感受野特征图更大的权重，若更应注意大目标则反之。此外，在二维 DCT 变换中频率分量的选择同上一节。

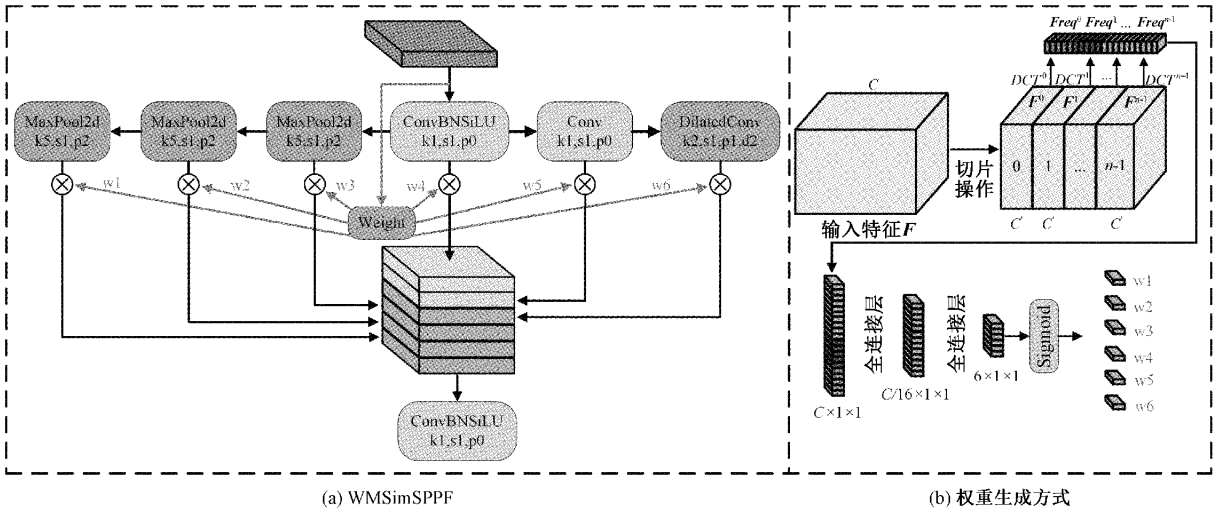


图 4　WMSimSPPF 结构

2.4　跨层注意力融合模块

在目标检测任务当中，网络的浅层特征感受野较小，所提取的特征与输入比较相近，包括更多如颜色、纹理、棱角和边缘等细粒度特征，这些特征更利于网络对于小目标的定位。而网络的深层特征经过多层卷积的运算包含更多的是抽象的语义信息，其中小目标的细节特征几乎被消除，这也是许多目标检测模型对小目标检测效果不理想的原因。基于以上原因，本文提出的改进模型在 YOLOX-S 的 neck 中进行跨层特征融合，提出一种跨层注意力融合模块(cross-layer attention fusion module，CAFM)，将浅层特征与深层特征相融合，并使用注意力机制减少特征融合的冗余。

如图 5 所示，使用 backbone 中 d2、d3 和 d4 层的输出特征，与 neck 中 PAN 的 3 个输出特征 p1、p2 和 p3 分别进行浅层特征与深层特征的跨层特征融合。与常规的跨层融合不同，直接将 backbone 中的特征与 PAN 的输出特征进行融合，而不再经过 FPN 和 PAN 的处理。相对于后者，通过直接与 PAN 的输出特征进行融合的方式，可以在融合深浅层特征的同时最小程度的增加模型的参数量，避免网络的过拟合；其次在网络的 backbone 部分，由于特征只经过了较少的处理，所保留小目标的细粒度特征较多，若将其在 FPN 和 PAN 中进行融合并经过过多的处理，会不可避免的造成小目标细粒度特征的丢失。

CAFM 的结构如图 6(a)所示，其中 $\boldsymbol{X} \in \mathbb{R}^{C\times H\times W}$ 代表

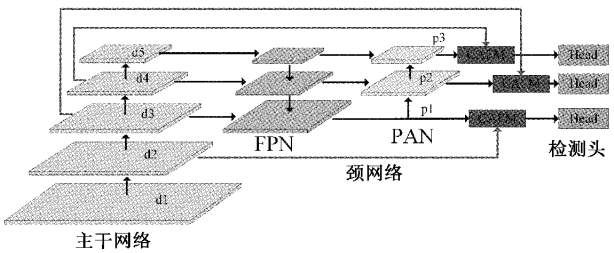


图 5　跨层融合结构

backbone 输出的浅层特征，$\boldsymbol{Y} \in \mathbb{R}^{2C\times\frac{H}{2}\times\frac{W}{2}}$ 代表 PAN 输出的深层特征。受文献[14]启发，使用 SPD(space-to-depth)层进行下采样，如图 6(b)所示。SPD 层将特征图 $\boldsymbol{X}$ 按相隔特征点进行切分，生成 4 个子特征图，其中每个子特征图的大小都为 $(C\times\frac{H}{2}\times\frac{W}{2})$，最后进行拼接操作生成下采样完成的特征图 $\boldsymbol{X}' \in \mathbb{R}^{4C\times\frac{H}{2}\times\frac{W}{2}}$。相比于直接使用池化等下采样方式，SPD 层将空间的细粒度特征转换为通道深度特征，并未直接将细粒度特征抹去。在 SPD 层之后使用一个非跨步(即步长为 1，核为 1)的动态卷积(dynamic convolution，DyConv)[15]对 $\boldsymbol{X}'$ 进行处理。动态卷积的结构如图 6(c)所示，其先对于输入的特征图进行注意力计算，生成 n 个注意力权重，然后对 n 个卷积核参数进行线性求和，将求和后的卷积核作为动态卷积的卷积核进行运算。相对于传统卷积，动态卷积可随着输入的变化而变

化。在 SPD 层之后使用动态卷积,可以通过较少的操作更有效的提取小目标特征。此外,相对于步长大于 1 的跨度会对特征信息造成无差别的损失,使用非跨步的方式更有助于网络保留小目标的细粒度信息。最后,CAFM 使用空间注意力对两个特征图进行注意力计算,减少特征融合过程中特征信息的冗余。

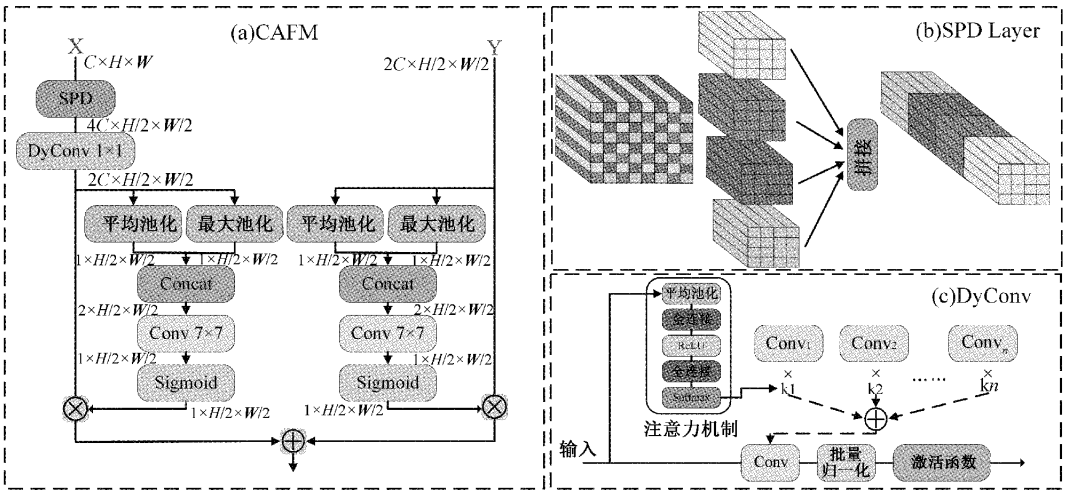


图 6 CAFM 整体结构

2.5 EIoU 损失

评价检测器是否可以准确的定位到目标物,通常用交异比(intersection over union,IoU)指标来评估,它代表着预测框与边界框的重合程度。而 YOLOX-S 的框回归损失是通过每个真实框所对应特征点预测出的预测框与该真实框计算广义交异比(generalized intersection over union,GIoU)[16]损失得到。但 GIoU 存在着一些问题:当预测框和真实框宽高相同且处在同一水平线时,GIoU 就退化成了 IoU 从而导致预测框回归不准确。

由于小目标对边界框的扰动较为敏感,本文考虑使用 EIoU 损失替换原模型的 GIoU 损失。EIoU 损失的计算方式如式(6)、(7)所示。

$$EIoU = IoU - \frac{\rho^2(b,b^{gt})}{c^2} - \frac{\rho^2(w,w^{gt})}{C_w^2} - \frac{\rho^2(h,h^{gt})}{C_h^2} \tag{6}$$

$$Loss_{EIoU} = 1 - EIoU \tag{7}$$

其中,c 代表预测框和真实框最小外接矩形的对角线长度,C_w 和 C_h 为最小外接矩形的宽度和长度,b 和 b^{gt} 分别为预测框和真实框的中心点,ρ 代表欧氏距离,w、w^{gt}、h 和 h^{gt} 分别代表预测框和真实框的宽度和长度。相比于 GIoU,EIoU 额外考虑了中心点距离和长宽比,可以使预测框定位的更加准确,以用来缓解小目标定位难的问题;同时,由于额外添加了中心点距离和长宽比损失,使得损失函数收敛速度更快。

3 实验结果及分析

3.1 实验环境和参数设置

本文当中模型训练和性能测评硬件配置为:Intel Core i7-12700KF(3.60 GHz),内存为 32 GB,GPU 型号为 NIVDIA RTX A4000,显存为 16 GB;软件环境为:Windows 10,python 3.8,pytorch 1.9,使用 CUDA 框架并行加速运算,CUDA 版本为 11.0。为了模型性能的公平比较,实验中统一使用 Adam 优化器,动量大小设为 0.937,学习率采用余弦退火算法,初始学习率为 1×10^{-3},最小学习率为初始学习率的 0.01 倍,batch 大小设为 16,epoch 设为 300。

3.2 实验数据集

当前,在遥感目标检测领域所提出的数据集有很多。考虑到数据集中类别数量、小目标丰富度,本文选择使用 DIOR[17]和 RSOD[18]作为本文的实验数据集。

DIOR 是一个用于光学遥感检测的大规模数据集,包含如高速服务区(C1)、高速收费站(C2)、飞机(C3)、机场(C4)、棒球场(C5)、篮球场(C6)、桥(C7)、烟囱(C8)、水坝(C9)、高尔夫球场(C10)、田径场(C11)、港口(C12)、立交桥(C13)、船(C14)、体育场(C15)、储罐(C16)、网球场(C17)、火车站(C18)、车辆(C19)和风车(C20)共 20 个种类。数据集共 23 463 张图像和 192 472 个实例,其中将 5 862 张作为训练集,5 863 张作为验证集,11 738 张作为测试集。RSOD 是一个开放的遥感数据集,包括飞机、油箱、操场和立交桥共 4 个种类。数据集共 976 张图片和 6 950 个实例,按照 8∶2 的比例随机分为训练集和测试集。

为证明实验所用数据集的可靠性,将数据集中目标框的长宽占比进行统计,如图 7 所示,其中图 7(a)为 DIOR 数据集长宽比,图 7(b)为 RSOD 数据集长宽比。所提及的两个数据中都存在着大量长宽比小于 10%的目标框,符合遥感小目标的实验要求。

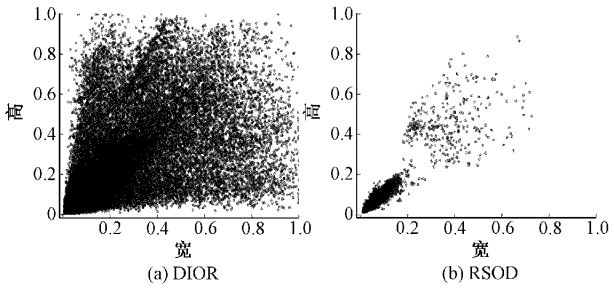


图 7　数据集长宽比

3.3　实验评价指标

本文采用目标检测模型常用评价指标：平均精度(average precision，AP)、均值平均精度(mean average precision，mAP)、检测速度(frames per second，FPS)和模型参数量(params)作为模型的评价指标。此外为了具体展现本文提出模型对于小目标的检测效果，COCO 评价指标中的 APs(average precision of small)也将作为评价指标之一。

3.4　消融实验

为验证本文所提各个改进方法的有效性，在 DIOR 数据集上设计了一组消融实验，实验结果如表 1 所示，“√”表示添加了该方法。本文所提的所有方法共使模型 mAP 提升了 3.07%，APs 提升了 2.4%，说明了在使用了全部改进方法的情况下，可显著提升模型对于遥感目标，尤其是小目标的检测能力。此外，在使用了所有方法后，参数量相较于原模型只增加了 1.01 M，FPS 只下降了 24 fps，只使用了较少的开销换取了较大的提升。从单个改进方法来看，使模型性能提升最大的是 CAFM，使 mAP 提升了 1.7%，APs 提升了 2%，但同时也使模型参数量增加了 0.37 M；其次是 WMSimSPPF 和 CBAM-MSCA，两者分别使模型 mAP 提升了 1.5%和 1.48%，APs 提升了 1.6%和 1.4%，模型参数量分别增加了 0.61 M 和 0.05 M；最后是 EIoU 损失的引入，其并不会增加模型的参数量，但同时也只带来了较小的提升，使 mAP 提升了 0.37%，APs 提升了 1.4%。此外，当所有改进方式逐个累加时，均能带来模型性能上的提高。

表 1　消融实验

CBAM-MSCA	WMSimSPPF	CAFM	EIoU	mAP/%	APs/%	参数量/M	检测速度/fps
—	—	—	—	70.91	11.2	8.93	117.9
√	—	—	—	72.39	12.6	8.98	115.2
√	√	—	—	72.60	12.9	9.59	101.0
√	√	√	—	73.62	13.5	9.94	93.6
√	√	√	√	73.98	13.6	9.94	93.6
—	√	—	—	72.41	12.8	9.54	105.2
—	—	√	—	72.61	13.2	9.29	108.7
—	—	—	√	71.28	12.6	8.93	117.9

从总体来看，本文改进模型在提高 mAP 和 APs 方面均有一定的优势。同时由于各个改进方法的加入，势必会提升模型的复杂度，在参数量和检测速度方面会稍逊色于原模型。但少许的参数量和检测速度的增加并不影响模型的轻量化和实时性，以少量开销换取检测精度的提升是极具性价比的。

3.5　对比实验

1) 注意力机制模型对比

为进一步验证改进 CBAM-MSCA 的有效性，本文以 YOLOX-S 为基线模型，做了一组注意力模块的横向对比实验，分别在模型相同位置加入 CMAM-MSCA、CBAM、坐标注意机制(coordinate attention，CA)[19]、选择卷积核(selective kernel，SK)[20]、MSCAM 和 CAM-MSCA 注意力模块在 DIOR 数据集上进行对比，实验结果如表 2 所示。由表 2 可知，不同的注意力模块均能带来性能的提升，而模型使用 CBAM-MSCA 在精度方面达到了最高。在模型参数量方面，CBAM-MSCA 的加入只使模型参数达到了 8.98 M，仅比 CA 高出 0.01 M，低于 CBAM 的 9.17 M。这是因为在二维 DCT 变换中使用的是预设的频率分量，因此在使用多频段分量进行特征压缩时不会引入额外的参数量。此外，相比于 MSCAM，模型使用 CAM-MSCA 的 mAP 值也要高出 0.37%。

表 2　注意力模块对比实验

注意力模型	mAP/%	APs/%	参数量/M
CBAM-MSCA	72.39	12.6	8.98
CBAM	71.68	12.0	9.17
CA	71.59	12.4	8.97
SK	71.05	11.5	38.0
MSCAM	71.87	12.2	8.98
CAM-MSCA	72.24	12.2	8.98

为了增加各个注意力模型性能的可解释性，使用热力图对上述 6 种注意力模型的效果在一张小目标遥感图像

上进行可视化分析,可视化结果如图 8 所示。其中 CBAM-MSCA 无论从热力图响应强度还是响应紧密度都达到了最好。其余注意力模型均存在一定的热源散漫不集中的问题,其中 SK 注意力模型受背景噪声影响较大,形成了较大的热源散漫和位置偏移。综上,通过二维 DCT 将频域信息引入到 CBAM 当中可更好的过滤背景噪声,能提升小目标检测的精度。同时,由于二维 DCT 并没有带来参数量,在改进后会一定程度减小模型参数。

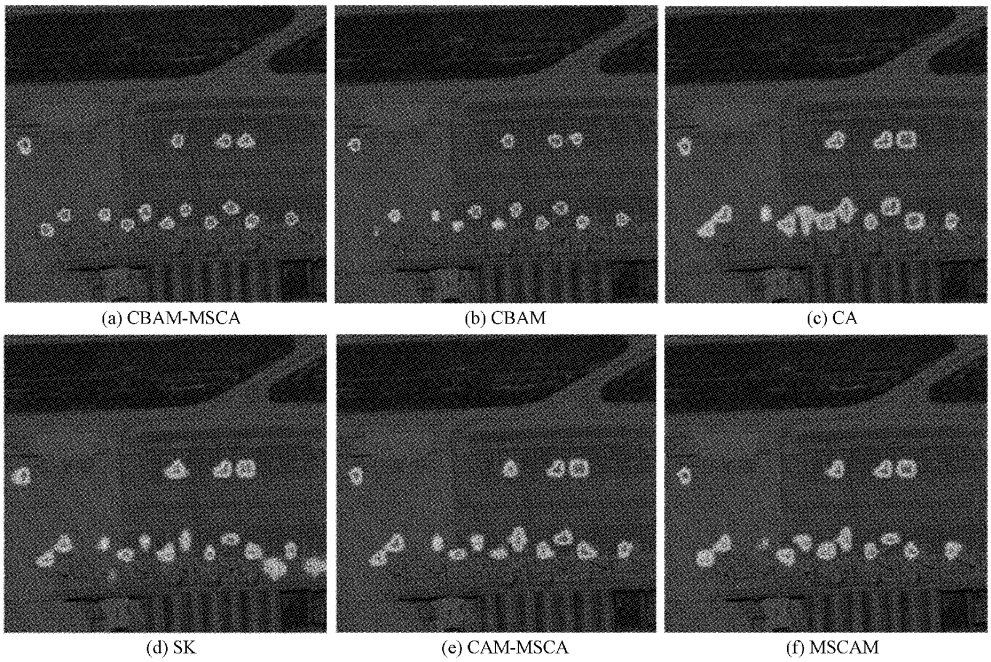


图 8 热力图可视化结果

2) 不同空间金字塔池化方式对比

为了探究 WMSimSPPF 的优越性,分别使用带有 SPP、SimSPPF 和 WMSimSPPF 的 YOLOX-S 模型进行了一组对比实验,实验结果如表 3 所示。其中使用 WMSimSPPF 的 YOLOX-S 的检测精度达到了最好,相对于使用 SPP 和使用 SimSPPF,mAP 值分别提高了 1.5% 和 0.38%,APs 值分别提高了 1.6% 和 0.7%。因此,通过在空间金字塔池化模块中构建额外的小感受野和自适应加权的方式,能够加强模型的多尺度感知能力和自适应尺度强化能力,可以有效提升模型对于小目标检测的精度。此外,因为添加了额外的小感受野和自适应权重,少许参数量的增多是不可避免的,但整体不影响模型的轻量化。

表 3 空间金字塔池化模块对比

模型	mAP/%	APs/%	参数量/M
YOLOX-S w/SPP	70.91	11.2	8.93
YOLOX-S w/SimSPPF	72.30	12.1	8.93
YOLOX-S w/WMSimSPPF	72.41	12.8	9.54

3) 不同跨层融合方式对比

为了对比不同融合方式的性能差别,在本文所提跨层特征融合处分别使用如图 9 所示的 3 种不同的融合方式,其中图 9(a)代表 CAFM,图 9(b)代表不使用动态卷积的 CAFM,图 9(c)代表使用 3×3 卷积进行下采样融合的方式,对比实验结果如表 4 所示。其中使用 CAFM 进行跨层融合得到的精度最高,相对于不使用动态卷积的 CAFM,mAP 和 APs 分别提高了 0.21% 和 0.6%;相对于使用常规 3×3 卷积进行下采样并融合方式,mAP 和 APs 分别提高了 0.65% 和 0.6%。在模型的参数量方面,模型使用 CAFM 的参数量也要小于使用常规 3×3 卷积进行融合方式。综上,相对于在传统跨层融合中使用 3×3 卷积进行下采样的方式,CAFM 在精度和参数量方面均有一定优势。

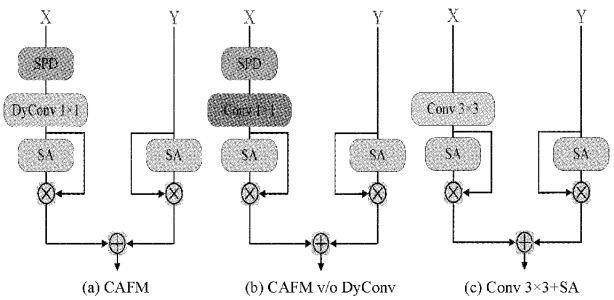


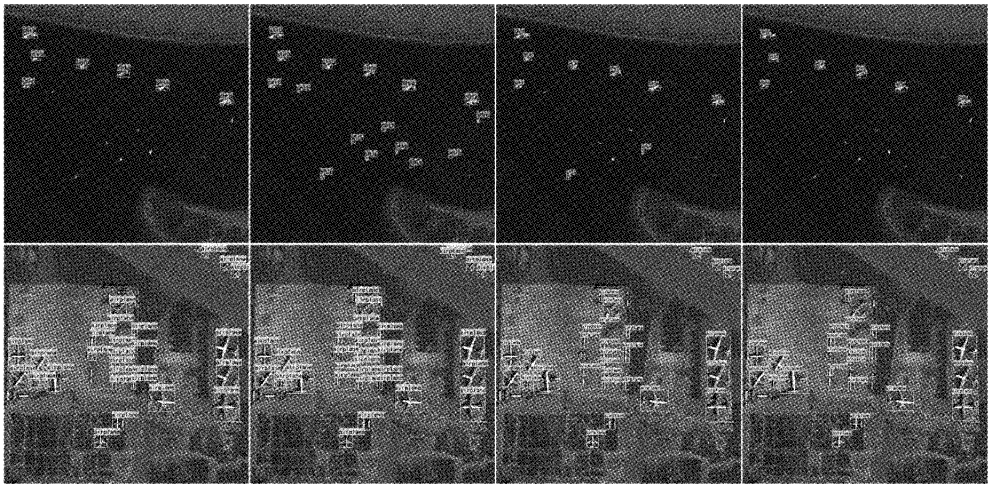
图 9 3 种融合方式

表 4 不同融合方式对比

融合方式	mAP/%	APs/%	参数量/M
CAFM	72.61	13.2	9.29
CAFM w/o DyConv	72.40	12.6	9.01
Conv3×3+SA	71.96	12.6	10.49

4) 模型检测效果对比

为了直接的体现本文改进模型对遥感小目标的检测能力,使用基线模型(baseline)、本文模型、YOLOv5-S 和 Faster-RCNN[21]进行检测效果对比,检测效果如图 10(a)~(d)所示。相对于另外 3 种常见模型在检测遥感小目标过程中存在着大量的漏检情况,本文改进模型在很大程度上改善了这一情况。



(a) 基线模型　(b) 本文模型　(c) YOLOv5-S　(d) Faxter-RCNN

图 10　本文改进模型与基线模型以及最常用模型的检测效果对比。

5) 与其他目标检测模型对比

最后,为了证明本文改进模型相较于当前主流目标检测模型的优越性以及在不同数据集当中的泛用性,分别在 RSOD 数据集和 DIOR 数据集上,与 YOLOv7[22]、YOLOv6、YOLOX-S、YOLOv5-S、YOLOv4[23]、Faster-RCNN、M2Det[24]和 FCOS[25]共 8 种目标检测模型进行性能的对比,对比结果如表 5 和 6 所示。由表中对比可知,本文模型在各项评价指标中,均处于较高水平。其中相对于目前最新目标检测模型 YOLOv7 和 YOLOv6-S 在看,在 RSOD 数据集上,相对于 YOLOv6-S 和 YOLOv7,本文模型 mAP 值要高出 2.65% 和 0.05%,APs 值分别高出 3.9%和 0.7%;在 DIOR 数据集上,本文模型 AP 值虽低出 YOLOv7 的 APs 值 0.1%,但在 mAP 值、参数量以及检测速度方面本文模型均有优势。相较于 YOLOv6-S,本文模型的精度指标均要高出一定的值。最后,在检测速度方面本文模型可达到 93.6 fps,满足实时检测的要求。

表 5　RSOD 数据集上各个模型性能

模型	AP/%				mAP/%	APs/%	参数量/M	检测速度/fps
	飞机	油桶	立交桥	操场				
YOLOv7	97.36	98.42	84.62	100	95.10	41.7	36.90	44.1
YOLOv6-S	96.40	98.70	95.70	100	97.70	44.6	17.19	133.0
YOLOX-S	96.37	98.19	83.78	100	94.58	40.2	8.93	117.9
YOLOv5-S	95.74	98.18	80.32	99.28	93.38	37.3	7.07	121.9
YOLOv4	93.51	97.26	87.17	99.83	94.44	27.6	63.90	62.2
Faster-RCNN	81.78	97.69	93.28	100	93.19	11.6	136.71	44.3
M2Det	88.11	97.87	94.01	100	95.00	14.7	86.50	50.2
FCOS	93.95	98.81	85.06	100	94.45	31.4	51.0	38.1
ours	96.60	98.18	96.22	100	97.75	45.3	9.94	93.6

表 6 DIOR 数据集上各个模型性能

模型	YOLOV7	YOLOV6-S	YOLOX-S	YOLOV5-S	YOLOV4	Faster-RCNN	M2Det	FCOS	ours
C1	60.43	69.5	66.37	62.25	67.23	60.23	66.44	70.85	63.21
C2	61.23	61.5	63.29	62.98	60.89	52.18	51.09	63.86	68.17
C3	90.57	89.6	83.49	86.55	84.03	58.29	72.25	87.91	90.73
C4	77.85	83.8	78.64	69.22	79.96	76.00	68.26	77.58	77.50
C5	76.68	82.2	74.94	81.45	80.16	71.00	69.53	80.87	80.51
C6	91.47	90.4	88.58	89.61	87.60	85.92	86.37	89.77	90.52
C7	46.23	42.8	41.68	40.15	38.46	29.65	28.63	43.89	47.68
C8	78.31	77.4	76.83	77.99	79.41	76.27	77.39	80.44	77.90
C9	58.88	59.4	63.78	50.34	60.02	53.03	56.48	57.64	66.07
C10	78.79	80.4	71.49	68.86	75.30	79.32	72.50	77.97	79.30
C11	80.38	78.1	74.22	75.93	76.10	64.86	65.65	74.75	80.64
C12	65.64	64.4	63.19	59.46	62.21	54.80	55.06	62.68	64.06
C13	61.84	59.3	58.59	57.98	58.69	53.52	51.59	58.94	61.80
C14	91.04	90.8	89.82	89.30	86.51	27.13	52.34	87.09	89.87
C15	68.14	79.1	70.43	64.99	61.43	69.47	62.61	66.53	72.42
C16	79.98	76.1	72.93	75.24	72.03	33.91	44.43	73.00	78.31
C17	90.31	90.2	85.89	89.57	88.29	80.82	81.03	90.67	90.54
C18	60.39	59.7	62.43	55.31	58.22	55.69	43.11	57.57	63.88
C19	56.03	50.7	52.74	52.23	48.56	17.04	28.17	49.01	55.66
C20	81.03	78.7	78.79	75.10	77.48	59.65	65.77	80.35	80.67
mAP/%	72.76	73.2	70.91	69.23	70.13	57.94	59.94	71.57	73.98
APs/%	13.70	13.50	11.20	11.70	9.10	1.30	2.10	11.80	13.60

4 结 论

针对于遥感图像中小目标难以检测的问题,本文对 YOLOX-S 目标检测模型进行改进。该方法首先通过二维 DCT 变换对 CBAM 注意力模块进行改进,将频率信息引入到注意力当中,使模型能够更好的过滤背景噪声;在 SimSPP 的基础上通过增加小感受野的特征图和自适应权重构建出一种加权多重感受野空间金字塔池化模块,加强模块的多尺度尤其是小尺度的感知能力;对模型进行跨层融合,提出一种跨层注意力融合模块,使模型可以最大程度保留小目标特征;最后使用 EIoU 损失加强对小目标的定位。在 DIOR 和 RSOD 2 个遥感数据集上证明了本文改进模型的有效性和泛用性。但由于改进模块的加入,势必会增加一定的参数量和检测速度,未来模型需要朝着高精度和轻量化的方向研究。

参考文献

[1] ZOU Z, CHEN K, SHI Z, et al. Object detection in 20years: A survey[J]. Proceedings of the IEEE, 2023, 111(3): 257-276.

[2] 牛浩青,欧鸥,饶姗姗,等. 改进 YOLOv3 的遥感影像小目标检测方法[J]. 计算机工程与应用,2022,58(13):241-248.

[3] 汪鹏,郑文凤,史进,等. 基于 MFANet 和上下文特征融合的遥感影像目标检测[J]. 应用科学学报,2022,40(1):131-144.

[4] 唐建宇,唐春晖. 基于旋转框和注意力机制的遥感图像目标检测算法[J]. 电子测量技术,2021,44(13):114-120.

[5] WOO S, PARK J, LEE J Y, et al. Cbam: Convolutional block attention module[C]. Proceedings of the European Conference on Computer Vision(ECCV), 2018: 3-19. DOI: 10.1007/978-3-030-01234-2_1.

[6] ZHANG Y F, REN W, ZHANG Z, et al. Focal and efficient IoU loss for accurate bounding box regression[J]. Neurocomputing, 2022, 506: 146-157, DOI: 10.1016/j.neucom.2022.07.042.

[7] GE Z, LIU S, WANG F, et al. YOLOX: Exceeding YOLO series in 2021[EB/OL]. (2021-08-06)[2022-12-27]. https://arxiv.org/abs/2107.08430.

[8] WANG C Y, LIAO H Y M, WU Y H, et al. Cspnet: A new backbone that can enhance learning capability of cnn[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020: 390-391, DOI: 10.1109/cvprw50498.2020.00203.

[9] LIN T Y, DOLLAR P, GIRSHICK R, et al. Feature

pyramid networks for object detection[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 936-944, DOI: 10.1109/CVPR.2017.106.

[10] LI H, XIONG P, AN J, et al. Pyramid attention network for semantic segmentation[EB/OL]. (2018-11-25) [2022-12-27]. https://arxiv.org/abs/1805.10180.

[11] QIN Z, ZHANG P, WU F, et al. Fcanet: Frequency channel attention networks[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, October 10-17, 2021, Montreal, QC, Canada. New York: IEEE, 2021: 783-792, DOI: 10.1109/ICCV48922.2021.00082.

[12] HE K M, ZHANG X Y, REN S Q, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis is and Machine Intelligence, 2015, 37(9): 1904-1916.

[13] LI C, LI L, JIANG H, et al. YOLOv6: A single-stage object detection framework for industrial applications[EB/OL]. (2022-09-07)[2022-12-27]. https://arxiv.org/abs/2209.02976.

[14] SUNKARA R, LUO T. No more strided convolutions or pooling: A new CNN building block for low-resolution Images and small objects[EB/OL]. (2022-08-07) [2022-12-27]. https://arxiv.org/abs/2208.03641.

[15] CHEN Y P, DAI X Y, LIU M C, et al. Dynamic convolution: Attention over convolution kernels[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 11027-11036, DOI: 10.1109/CVPR42600.2020.01104.

[16] REZATOFIGHI H, TSOI N, GWAK J Y, et al. Generalized intersection over union: A metric and a loss for bounding box regression[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), June15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 658-666, DOI: 10.1109/CVPR.2019.00075.

[17] LI K, WAN G, CHENG G, et al. Object detection in optical remote sensing images: A survey and a new benchmark[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2020, 159(1): 296-307.

[18] LONG Y, GONG Y P, XIAO Z F, et al. Accurate object localization in remote sensing images based on convolutional neural networks[J]. IEEE Transactions on Geoscience and Remote Sensing, 2017, 55(5): 2486-2498.

[19] WEN W, XU C, WU C P, et al. Coordinating filters for faster deep neural networks[C]. 2017 IEEE International Conference on Computer Vision, October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 658-666. DOI: 10.1109/ICCV.2017.78.

[20] LI X, WANG W H, HU X L, et al. Selective kernel networks[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 510-519, DOI: 10.1109/CVPR.2019.00060.

[21] REN S Q, HE K M, 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.

[22] WANG C Y, BOCHKOVSKIY A, LIAO H Y M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[EB/OL]. (2022-07-06) [2022-12-27]. https://arxiv.org/abs/2207.02696.

[23] BOCHKOVSKIY A, WANG C-Y, LIAO H-Y M. Yolov4: Optimal speed and accuracy of object detection[EB/OL]. (2022-04-23)[2022-12-27]. https://arxiv.org/abs/2004.10934.

[24] ZHAO Q J, SHENG T, WANG Y T, et al. M2Det: A single shot object detector based on multi-level feature pyramid network[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 9259-9266.

[25] TIAN Z, SHEN C H, CHEN H, et al. FCOS: Fully convolutional one-stage object detection[C]. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Korea(South). New York: IEEE Press, 2019: 9626-963, DOI: 10.1109/ICCV.2019.00972.

作者简介

张绍文，硕士研究生，主要研究方向为图像处理，计算机视觉。

E-mail:zsw219219@163.com

史卫亚(通信作者)，博士，副教授，硕士生导师，主要研究方向为深度学习、图像处理。

E-mail:swymail@126.com

张世强，硕士研究生，主要研究方向为模式识别，图像处理。

E-mail:2843056239@qq.com

王甜甜，硕士研究生，主要研究方向为图像处理，计算机视觉。

E-mail:wtt1569230632@163.com