文章编号:1671-6833(2024)03-0072-08

基于特征融合和混合注意力的小目标检测

魏明军[1,2], 王镆涵[1], 刘亚志[1,2], 李 辉[1]

(1.华北理工大学 人工智能学院,河北 唐山 063210;2.华北理工大学 河北省工业智能感知重点实验室,河北 唐山 063210)

摘 要:针对目标检测任务中小目标特征信息不足、检测率较低,且错、漏检率较高等缺点,提出一种基于多尺度特征融合以及混合注意力机制的 Tr-SSD 算法。首先,使用 Resnet50 残差网络作为 SSD 算法的骨干网络,增强 SSD 算法的特征提取能力;其次,设计了一种混合注意力机制并将其应用于网络的中尺度特征图中以增强特征图中的有效信息,并建立信息间的远距离依赖;最后,使用以 Transformer 为核心的网络层与替换骨干网络后的 SSD 算法形成 FPN 结构,融合不同尺度的特征信息,以更准确地对小目标进行定位。实验结果表明:Tr-SSD 算法在 PASCAL VOC 数据集、HRSID 数据集和 RSOD 遥感数据集上检测的 *mAP* 值分别达到 81.9%、87.5%和 88.4%,比 SSD 算法分别提高了 4.7 百分点、6.8 百分点和 9.2 百分点,且检测速度均满足实时检测的要求。

关键词:小目标检测; 注意力机制; 特征融合; 深度学习; 实时检测

中图分类号:TP391.4;TP18 **文献标志码**:A **doi**:10.13705/j.issn.1671-6833.2024.03.001

随着深度学习技术的飞速发展,目标检测技术已经在工业、交通、医疗、遥感、安全等各个领域取得了广泛应用,如安全保障、工业检测等。目前目标检测算法可以分为以 SSD 系列算法与 YOLO 系列算法为代表的单阶段目标检测算法和以 Faster R-CNN、FPN、Mask R-CNN 等算法为代表的两阶段目标检测算法两大类。其中,两阶段目标检测算法先在区域生成一个可能包含待检测物体的候选框,再进行样本分类,这类算法精度较高,但速度较低。单阶段目标检测算法只需对目标提取一次特征即可实现目标检测,这类算法速度较快,但算法精度较低。

尽管近年来两类算法均已取得较大发展,但仍受限于小目标检测领域。小目标检测是目标检测中的一个难点,所谓小目标是指目标成像尺寸较小,例如在 COCO 数据集中,尺寸小于 32×32 像素的目标通常被定义为小目标。小目标有特征不明显、携带信息少等特点,如何提高小目标的检测精度一直是研究的热点。Liu 等[1]在 2016 年提出的 SSD 算法直接回归目标类别与位置,在不同尺度的特征图上进行预测,但浅层特征图缺乏语义信息,深层特征图缺乏位置信息,导致小目标的检测准确率较低。Fu 等[2]在 2017 年提出的 DSSD 算法使用 Resnet101 骨干网络替换 SSD 中的 VGG16 骨干网络,分类回归之前引入了残差模块,在 SSD 的辅助卷积层后又添加了反卷积层。相较于 SSD 算法,DSSD 算法在小目标的检测精度上有了较大提高,但 DSSD 算法所使用的 Resnet101 网络结构过深,因此其检测速度较慢且模型结构灵活性差。Lin 等[3]提出了 FPN 算法,通过引入由深到浅的特征信息,融合不同尺度特征图的方式保留语义信息。Tan 等[4]提出了小目标检测算法 EfficientDet,使用混合缩放的方法来缩放模型,并使网络自适应更新权重,但 EfficientDet 的基准模型是通过谷歌丰富的计算资源搜索得到的,计算花销较大。Qiao 等[5]提出的小目标检测算法 DetectoRS 将 FPN 的输出结果重新输入到骨干网络中进行二次提取并且可以自适应选择合适的感受野,但切换空洞卷积操作耗时严重。Lim 等[6]提出了小目标检测算法 FA-SSD,在 SSD 算法基础上构建用于融合上下文特征的上下文信息模块,增强算法对特征图中上下文信息的利用能力,并且在算法中

收稿日期:2023-09-20;**修订日期**:2023-10-19

基金项目:科技部重点研发项目(2017YFE0135700);河北省高等学校科学技术研究项目(ZD2022102)

作者简介:魏明军(1969—),男,河北唐山人,华北理工大学教授,主要从事计算机视觉、入侵检测、机器学习、数据挖掘等方面的研究,E-mail:weimj@ncst.edu.cn。

引用本文:魏明军,王镆涵,刘亚志,等. 基于特征融合和混合注意力的小目标检测[J]. 郑州大学学报(工学版),2024,45(3):72-79.(WEI M J, WANG M H, LIU Y Z, et al. Small object detection based on feature fusion and mixed attention[J]. Journal of Zhengzhou University (Engineering Science),2024,45(3):72-79.)

引入残差注意力机制,使得算法可以着重聚焦于图像的一部分,而不是整个区域。相较于 SSD 算法,FA-SSD 算法在小目标的检测精度上有一定的提高,但 FA-SSD 算法丢失了浅层纹理信息,导致大、中目标的检测精度略有下降,并且算法结构较为复杂,检测速度较慢。Yang 等[7]提出了小目标检测算法 QueryDet,预测头在低分辨率的特征图上预测小目标的粗略位置,检测头采用粗略位置稀疏引导的高分辨率特征计算准确的检测结果。QueryDet 算法充分利用了高分辨率特征图,又避免了背景区域的无效计算,在提高检测精度的同时保证了检测速度,但即使预测到了小目标的大概位置,检测头也可能无法对其定位,同时存在大、中目标错误激活检测头的可能,导致检测头处理无用位置,增加冗余计算。

相较于 YOLO 系列算法,SSD 算法的结构较为简单,可改进性较强,并且 SSD 算法采用了不同尺度和长宽比的先验框,更适合小目标检测任务,因此本文在 SSD 算法的基础上进行改进。SSD 使用了 VGG16 作为骨干网络,但存在提取目标信息能力不足、算法运行速度较慢等问题。本文使用网络层数适中且特征提取能力较强的 Resnet50 网络作为骨干网络替换原本的 VGG16 网络,同时,借鉴 FPN 的设计思想,使用以 Resnet50 为骨干网络的 SSD 与以 Transformer 结构为核心的网络层综合设计出特征融合模块,通过融合不同尺度的特征图来解决特征信息丢失的问题,提高小目标检测率,并且设计出以 CBAM 注意力机制与 Selfattention 机制为基础的混合注意力模块嵌入到网络结构中,以此关注并调整目标特征中的有效信息并联系不同特征图中的上下文信息。

1　相关工作

1.1　SSD 算法

SSD 算法是一种单阶段的目标检测算法,通过卷积神经网络进行特征提取,使用不同的特征层进行检测输出。SSD 算法在进行目标检测任务时会生成 6 个不同尺度的特征图,其中高层特征图感受野较大,适合检测大尺度目标,但缺乏位置信息;低层特征图感受野较小,适合检测小尺度目标,但缺乏语义信息。SSD 算法结构如图 1 所示。

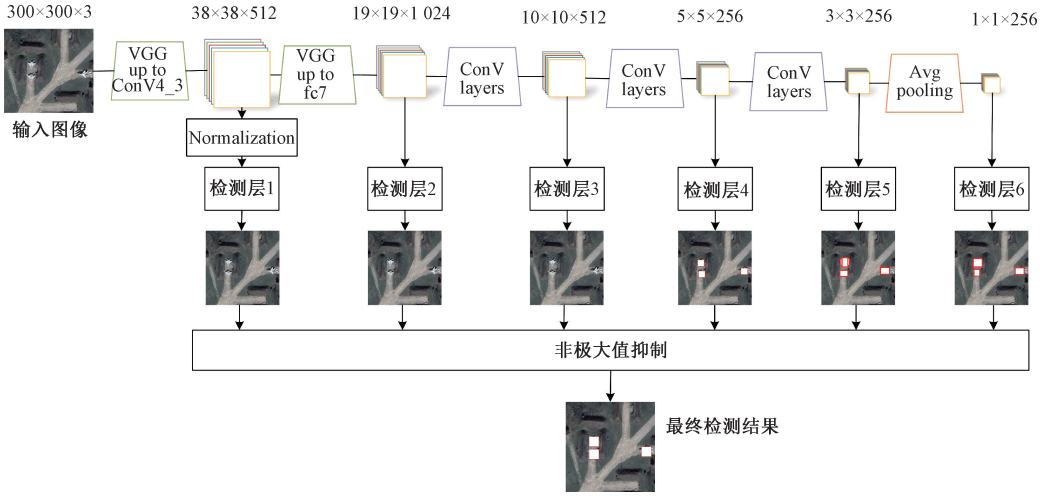


图 1　SSD 网络结构

Figure 1　SSD network structure

1.2　Transformer

Transformer 和 VIT(vision transformer)两种基于 Transformer 结构的神经网络模型分别用于文本序列和图像序列的建模。这两种模型的共同点是采用了 Transformer 结构,但是在细节上有所不同。Transformer 最初是作为一种用于自然语言处理的模型而被提出,其特点是能够准确地捕捉长文本序列中的依赖关系。与传统的递归神经网络(RNN)和卷积神经网络(CNN)相比,Transformer 具有更快的训练速度和更好的并行性能。其中,自注意力机制 Self-attention 是 Transformer 的核心,它通过计算每个词语与所有其他词语之间的相关性来得到特征表示。在计算机视觉领域,CNN 一直是最常用的算法。为了将 Transformer 应用于图像,VIT 被提出。

VIT 将图像看作是一个序列,并将每个像素点视为序列中的一个词语。VIT 能够处理高分辨率的图像,并且在一些图像分类任务中具有与传统 CNN 相当的性能。相比于传统的 CNN,Transformer 和 VIT 可以并行计算,因此二者在训练速度和效率上具有更大的优势。

2　Tr-SSD 算法

本文提出了一种基于多尺度特征融合以及混合

注意力机制的 Tr-SSD 算法,如图 2 所示。首先,该算法将 SSD 原本的骨干网络 VGG16 替换为 Resnet50,通过 Resnet50 网络提取输入图像中的位置信息与语义信息,相较于 VGG16,Resnet50 采用跨层链接,特征提取能力更强,能够有效提取小目标中的特征信息,并且在更深的网络层数时,较少发生梯度爆炸或消失等问题。其次,在 SSD 原有的前 3 个特征提取层(C1、C2、C3)前加入混合注意力模块(selective selfattention,SeSA),以强化特征图中的上下文信息,增强特征图中的关键信息。再次,算法加入了特征融合模块以丰富特征中的语义信息,增强网络对于小尺寸对象的特征提取能力。最后,将融合后的特征图送入检测头中,得到算法的最终检测结果,相较于 SSD 网络,Tr-SSD 算法在小目标检测任务中的表现更好,在大、中目标的检测任务上同样取得了理想效果。

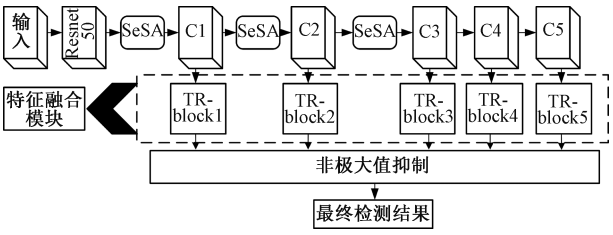


图 2 Tr-SSD 算法框架

Figure 2 Tr-SSD algorithm block diagram

2.1 混合注意力模块

由于小目标的尺寸较小,携带的特征信息也较为缺乏,在特征提取的过程中,小目标所携带的特征信息易受到噪声的干扰,导致小目标的检测率较低,因此,为了将目标与场景中的杂乱干扰进行区分,本文在 CBAM(convolutional block attention module)基础上提出了 SeSA 混合注意力机制。CBAM 是一种用于计算机视觉领域的注意力机制模块,主要用于增强算法对于通道与空间信息的关注,但其有两点不足:①没有捕获不同尺度的空间信息来丰富特征空间;②空间注意力仅仅考虑了局部区域的信息,而无法建立远距离的依赖。

SeSA 由通道注意力、空间注意力和多头自注意力组成,是一种混合注意力机制,旨在增强目标检测网络中的特征信息和杂乱背景下的抗干扰能力,如图 3 所示。

首先,SeSA 使用通道注意力对输入特征进行通道维度上的细化,即对输入的特征 $\boldsymbol{F}$ 同时进行最大池化和平均池化,并使用共享全连接层和 sigmoid 归一化处理得到通道注意力权重 $\boldsymbol{w}_c$:

$$\boldsymbol{w}_c = \sigma(\mathrm{MLP}(\mathrm{AvgPool}(\boldsymbol{F})) + \mathrm{MLP}(\mathrm{MaxPool}(\boldsymbol{F}))); \quad (1)$$

$$\boldsymbol{F}_c = \boldsymbol{w}_c \cdot \boldsymbol{F}。 \quad (2)$$

式中:$\sigma(\cdot)$表示 sigmoid 操作。

其次,SeSA 使用空间注意力对特征 $\boldsymbol{F}_c$ 进行空间维度的细化。空间细化时先进行最大池化和平均池化,再进行 7×7 卷积与 sigmoid 的操作:

$$\boldsymbol{w}_s = \sigma(f^{7\times7}(\mathrm{AvgPool}(\boldsymbol{F}_c),\mathrm{MaxPool}(\boldsymbol{F}_c))); \quad (3)$$

$$\boldsymbol{F}_s = \boldsymbol{w}_s \cdot \boldsymbol{F}_c。 \quad (4)$$

式中:$f^{7\times7}$表示 7×7 卷积;$\boldsymbol{w}_s$ 为生成的空间注意力。

最后,SeSA 使用多头自注意力自适应学习特征间的远距离依赖关系。具体来说,假设 $\boldsymbol{F}_s$ 的维度为 $[B,H,W,C]$,其中 B 表示输入的样本数量;H 和 W 分别表示特征图的高和宽;C 表示特征图的通道数。进行多头自注意力时,通过 3 个不同的权重矩阵对输入特征图进行线性变换,得到 3 个新的特征图,分别表示查询 $\boldsymbol{Q}$、键 $\boldsymbol{K}$ 和值 $\boldsymbol{V}$:

$$\begin{cases} \boldsymbol{Q} = \boldsymbol{W}_q \boldsymbol{F}_s; \\ \boldsymbol{K} = \boldsymbol{W}_k \boldsymbol{F}_s; \\ \boldsymbol{V} = \boldsymbol{W}_v \boldsymbol{F}_s。 \end{cases} \quad (5)$$

式中:$\boldsymbol{W}_q$、$\boldsymbol{W}_k$ 和 $\boldsymbol{W}_v$ 均为线性变换时的权重矩阵。

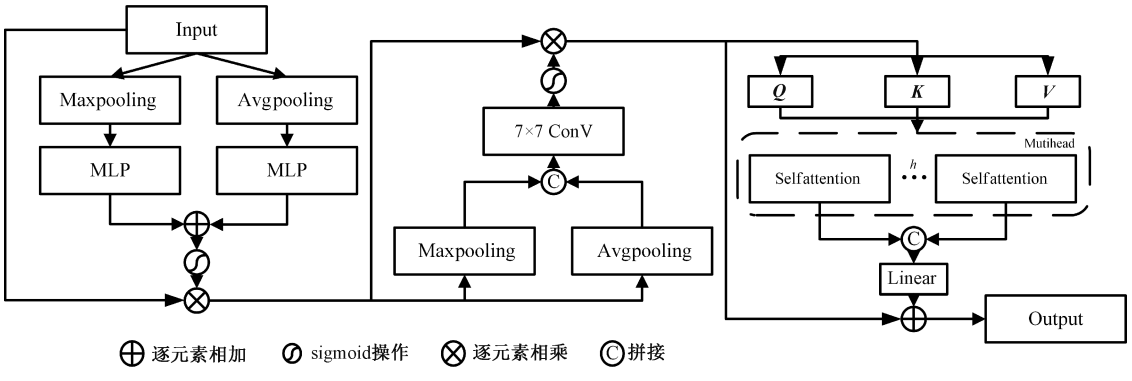


图 3 SeSA 混合注意力模块

Figure 3 SeSA hybrid attention module

变换后的特征图维度为[B,H,W,d_k],其中 d_k 为每个头部的特征向量维度。将变换后的特征图沿着最后一个维度分成 h 个头部。这里的 h 是一个超参数,用于控制切分后的头部数量。切分后的特征图维度为[B,H,W,h,d_k]。对于每个头部,分别计算查询和键的相似度矩阵,对相似度矩阵进行 softmax 归一化,得到每个键对查询的注意力权重。注意力加权后的特征图维度为[B,H,W,h,d_k],将每个头部的注意力加权结果进行拼接,并通过线性变换将拼接后的结果映射回原始特征向量的维度,得到最终的注意力加权结果。接下来将注意力加权后的特征图与输入特征图进行残差连接,最终得到输出特征图。处理后的特征图维度与输入特征图相同,均为[B,H,W,C]。以上操作可以表示如下:

$$\boldsymbol{O} = \mathrm{Linear}(\mathrm{Concat}(\boldsymbol{H}_i, \boldsymbol{H}_{i+1}, \cdots, \boldsymbol{H}_h)) + \boldsymbol{F}_s; \quad (6)$$

$$\boldsymbol{H}_i = \mathrm{Selfattention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V})。 \quad (7)$$

式中:Concat 表示拼接操作;$\boldsymbol{H}_i$ 表示第 i 个自注意力生成的特征图。

相较于 CBAM,SeSA 能够在加强远距离的信息依赖的同时,考虑不同尺度的空间信息,从而更好地挖掘特征信息,弥补了 CBAM 的缺点,更加适合小目标检测任务。

2.2　特征融合模块

SSD 算法使用多尺度的特征图进行预测,但浅层特征图中缺乏语义信息,导致 SSD 算法在进行小目标检测任务时表现较差,而深层特征图中包含大量的语义信息,因此,为了更好地利用深层特征图中的语义信息,提高算法对于小目标的检测率,本文提出一种基于 Transformer 网络层的特征融合模块,在中高层的较小尺寸特征图中构建 TrFPN 结构,融合不同尺度的特征信息。TrFPN 块结构如图 4 所示,其中 $\boldsymbol{C}_i(i=1, 2, 3, 4, 5)$代表特征提取层。TrFPN 块由一个上采样模块和多个 Transformer 层组成,通过以多头自注意力机制为核心的特征融合操作,保留了不同尺度下特征图的信息,减少了特征提取时的损失,并且有效联系了全局的特征信息,提高了本文算法对小目标的检测能力。

在 TrFPN 中,对于第 i 个 TrFPN 块,其输入特征图为 $\boldsymbol{F}_i^{(\mathrm{in})}$,输出特征图为 $\boldsymbol{F}_i^{(\mathrm{out})}$,其中 $\boldsymbol{F}_i^{(\mathrm{in})}$ 是由上一层输出特征图上采样得到的。第 i 个 TrFPN 块的特征融合公式为

$$\boldsymbol{F}_i^{(\mathrm{out})} = \begin{cases} LN(FFN(MHA(\boldsymbol{C}_5, \boldsymbol{C}_5, \boldsymbol{C}_5))); \\ LN(FFN(MHA(\boldsymbol{F}_i^{(\mathrm{in})}, \boldsymbol{F}_i^{(\mathrm{in})}, \boldsymbol{F}_i^{(\mathrm{in})})) + \\ \quad \boldsymbol{C}_i), i = 1,2,3,4。 \end{cases} \quad (5)$$

式中:*MHA* 表示多头自注意力机制;*FFN* 表示前馈神经网络;*LN* 表示层归一化。

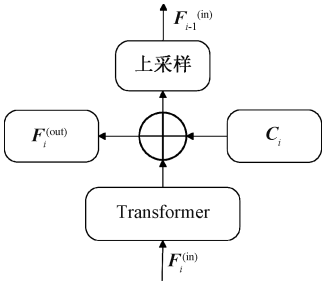


图 4　TrFPN 块结构图

Figure 4　TrFPN block structure diagram

2.3　算法流程

为了更直观地展现所提出算法的算法流程与结构,Tr-SSD 算法流程图如图 5 所示。

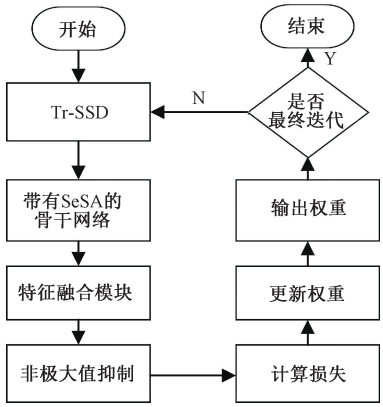


图 5　Tr-SSD 算法流程图

Figure 5　Tr-SSD Algorithm flow chart

算法 1　Tr-SSD 算法。

输入:X 为图像数据、L 为 Resnet50、T 为混合注意力模块、P 为特征融合模块、*epoch* 为模型训练次数;

输出:训练损失 *train_loss*、验证损失 *val_loss*、权重文件 *weight*。

① 将 X 划分为训练集 X_{train}、验证集 X_{val}、测试集 X_{t};
② 初始化迭代次数为 1;
③ *epoch* = 300;
④ for i = 1 to *epoch* do;
⑤ for $\boldsymbol{x}$,$\boldsymbol{y}$ in X_{train};
⑥ L 模块对输入 $\boldsymbol{x}$ 提取初始特征,得到 $\boldsymbol{x}'$;
⑦ T 模块对 $\boldsymbol{x}'$进行加强特征提取得到 $\boldsymbol{x}''$;
⑧ P 模块对 $\boldsymbol{x}''$进行特征融合,得到 $\boldsymbol{p}$;
⑨ 计算 *train_loss*;
⑩ for $\boldsymbol{t}$,$\boldsymbol{u}$ in X_{val};
⑪ L 模块对输入 $\boldsymbol{p}$ 提取初始特征,得到 $\boldsymbol{t}'$;
⑫ T 模块对 $\boldsymbol{t}'$进行加强特征提取得到 $\boldsymbol{t}''$;
⑬ P 模块对 $\boldsymbol{t}''$进行特征融合,得到 $\boldsymbol{o}$;

⑭ 计算 *val_loss*;
⑮ 获得 300 个权重;
⑯ 选择最优权重对测试集进行评估;
⑰ 获得平均精度均值 *mAP*;
⑱ 结束。

3 实验结果与分析

3.1 数据集与评价指标

本文在目标检测领域公共数据集 PASCAL VOC 07+12、小目标检测领域公开数据集 HRSID、遥感领域数据集 RSOD、自制吸烟数据集上进行实验。通过 PASCAL VOC 数据集测试算法的整体识别能力。PASCAL VOC 07+12 数据集包含背景在内共 21 个类别,训练集为包含 PASCAL VOC 07+12 的训练集,测试集为 VOC2007 测试集。

为了验证本文算法对于小目标的测试能力,选取 RSOD 遥感数据集和小目标检测领域数据集 HRSID 进行实验。RSOD 是一个包括飞机、油箱、游乐场和立交桥 4 种类别的遥感图像数据集。HRSID 是 2020 年电子科技大学发布的高分辨率 SAR 图像中用于船舶检测的通用小目标数据集,该数据集共包含 5 604 张高分辨率 SAR 图像和 16 951 个舰船目标。HRSID 和 RSOD 数据集的格式均为 PASCAL VOC,训练集与测试集的比例均为 8∶2。

为了进一步验证算法在实际应用中的能力,本文选取吸烟场景进行测试,同时为扩展小目标检测算法在实际场景中的应用做准备,如进行森林、工地、商场等禁烟场所的吸烟检测系统的研究。通过影视作品截图、网络爬虫采集、实地拍照 3 种方式自制了包含吸烟 1 种类别的吸烟数据集,以测试算法的泛化能力。自制吸烟数据集共 4 445 张图像,包含多种场景下的吸烟动作,数据集格式为 PASCAL VOC,使用 labelImg 软件对图片进行标注,其中训练集与测试集比例为 8∶2,自制吸烟数据集示例图片如图 6 所示。

图 6 自制吸烟数据集示例

Figure 6 Sample homemade smoking data set

本文采用 *mAP*(mean average precision)和 *FPS*(frames per second)作为评价指标,*mAP* 表示所有类别 *AP* 的平均值。

$$AP = \int_0^1 p(r)\,\mathrm{d}r; \tag{6}$$

$$mAP = \frac{\sum_{i=1}^{K} AP_i}{K}。 \tag{7}$$

式中:$p(r)$表示以召回率为横轴、准确率为纵轴的曲线;K 表示类别数。

预测框与真实框的交并比 $IOU>0.5$ 时,预测结果是正确的,交并比计算公式如下:

$$IOU = \frac{S_{交}}{S_{并}}。 \tag{8}$$

3.2 实验设置

本文实验基于 linux 操作系统,实验框架 Pytorch,使用 SGD(stochastic gradient descent)进行学习率调整,初始学习率设置为 0.001,权值衰减为 0.05,*batch size* 设置为 16,迭代 300 次后得到最终的网络模型。实验均在型号为 NVIDIA RTX 3090Ti,显存为 24 GB 的显卡上进行。

3.3 PASCAL VOC 实验对比分析

PASCAL VOC 数据集数据量庞大且包括大、中、小目标,适合测试本文算法的综合目标检测能力。结果对比如表 1 所示。

表 1 PASCAL VOC 数据集定量实验结果对比

Table 1 Quantitative comparison of experimental results of PASCAL VOC data set

算法	骨干网络	输入尺寸/像素	*mAP*/%	*FPS*/(帧·s^{-1})
Faster RCNN	VGG16	1 000×600	72.8	8.6
Mask RCNN	Resnet50	1 000×600	76.4	11.5
YOLOv3[8]	Darknet53	416×416	79.3	39.0
YOLOv5[9]	CSPDarknet53	416×416	80.4	44.3
SSD	VGG16	300×300	77.2	44.3
		512×512	79.1	21.5
DSSD	Resnet-101	300×300	79.2	11.4
EfficientDet[4]	efficientNet	640×640	80.2	40.6
FA-SSD[6]	VGG16	300×300	78.1	30.0
MFFAMM[10]	VGG16	300×300	80.7	35.2
AD-SSD[11]	VGG16	300×300	78.4	54.1
DP-SSD[12]	VGG16	300×300	81.2	30.7
Tr-SSD	Resnet50	300×300	81.9	35.6
		512×512	84.8	21.0

由表 1 可知,当图片输入尺寸为 300×300 像素时,本文算法的 *mAP* 值可以达到 81.9%,检测速度为 35.6 帧/s。相较于两阶段算法 Faster RCNN、Mask RCNN,Tr-SSD 算法 *mAP* 分别提高了 9.1 百分

点、5.5 百分点。当图片输入尺寸为 300×300 像素时,相较于 DP-SSD 算法,Tr-SSD 算法 *mAP* 提升了 0.7 百分点。相较于 AD-SSD 算法,Tr-SSD 算法的网络结构相对较为复杂,*FPS* 的值较低,但仍能满足实时检测要求,且 Tr-SSD 算法的 *mAP* 值提高了 3.5 百分点,整体性能优于 AD-SSD 算法。

为了更好地验证本文提出的替换骨干网络、使用以 Transformer 为核心的网络层与 SSD 构建 FPN 结构、增添混合注意力机制三者的有效性,在 PASCAL VOC 07+12 数据集上分别单独加入以上改进点进行消融实验,消融实验结果如表 2 所示。

表 2　消融实验结果对比

Table 2　Comparison of ablation results

算法	Resnet50	TrFPN	SeSA	*mAP*/%
SSD				77.2
	√			78.1
	√	√		79.5
	√		√	79.3
	√	√	√	81.9

注:√表示模块生效。

由表 2 可知,相比较与 SSD 算法,将 SSD 骨干网络替换为 Resnet50 后 *mAP* 值提高了 0.9 百分点,证明了 Resnet50 具有更好的特征提取能力。将 SSD 骨干网络替换为 Resnet50 并添加 TrFPN 模块以后提高了 2.3 百分点,说明了 TrFPN 模块可以很好地融合有效特征信息,提高算法的检测精度。混合注意力机制的增添同样提高了网络检测精度,相较于原始 SSD 算法,增添混合注意力机制后的 SSD 算法 *mAP* 值提高了 2.1 百分点。所有改进点综合以后,*mAP* 由原始 SSD 算法的 77.2%上升到 81.9%,提高了 4.7 百分点,证明了替换骨干网络、使用以 Transformer 为核心的网络层与 SSD 构建 FPN 结构、增添混合注意力机制这三者的有效性。

3.4　HRSID 数据集与 RSOD 数据集实验对比分析

本文通过多种算法在 HRSID 数据集中得到的实验数据进行对比,测试本文算法对于小目标的检测能力,定量实验数据对比结果如表 3 所示。

由表 3 可知,相较于对比算法中的最优算法 Efficient-YOLO,Tr-SSD 算法的 *mAP* 值提升了 0.5 百分点,这是由于在检测过程中,Tr-SSD 算法所使用的混合注意力机制建立了特征间的全局联系,有效增强了小目标的特征信息,并且在特征融合的过程中,充分利用了特征的语义信息,有效提高了对于小目标的检测率。

表 3　HRSID 数据集定量实验对比

Table 3　Quantitative experimental comparison of HRSID data sets

算法	输入尺寸/像素	骨干网络	*mAP*/%
Faster RCNN	1 000×600	VGG16	80.8
SSD	300×300	VGG16	80.7
YOLOv5	416×416	CSPDarknet53	83.3
CenterNet[13]	512×512	Resnet18	84.6
FCOS[13]	800×1 024	Resnet50	84.9
Free_anchor[14]	512×512	Resnet50	86.4
ASAFE[15]	500×500	Resnet50	85.1
Efficient-YOLO[16]	800×800	MobileNetv3	87.0
Tr-SSD	300×300	Resnet50	87.5

RSOD 遥感数据集中的大部分目标为小目标,通过各类算法在 RSOD 遥感数据集中得到的实验数据进行对比,测试本文算法对于小目标的检测能力,定量实验数据对比结果如表 4 所示。

表 4　RSOD 遥感数据集定量实验对比

Table 4　Quantitative experimental comparison of RSOD remote sensing datasets

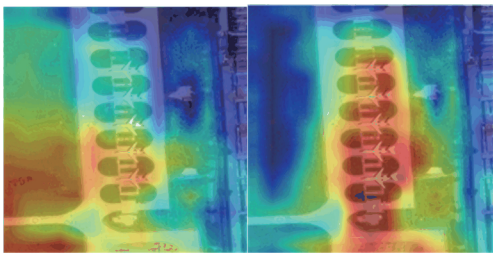
算法	骨干网络	*FPS*/(帧·s^{-1})	*mAP*/%
Faster RCNN	VGG16	18	81.7
SSD	VGG16	48	79.2
RSSD	VGG16	36	82.9
DSSD	Resnet-101	17	84.3
UAV-YOLO[17]	Darknet53	30	77.8
DC-SPP-YOLO[18]	DenseNet	33	76.6
Tr-SSD	Resnet50	56	88.4

从表 4 可以看出,Tr-SSD 算法在保持检测速度的同时检测精度达到 88.4%,优于其他算法。与对比算法中最优检测精度相比,*mAP* 提高了 4.1 百分点,证明了 Tr-SSD 算法能够充分利用小目标所携带的特征进行检测,使得检测精度相较于其他算法有所提升,且满足实时检测要求。

在 RSOD 数据集中进行热力图实验,结果如图 7 所示。图 7(a)为未添加混合注意力机制所得到的热力图,可以看到算法关注的位置并不精确。图 7(b)为添加混合注意力机制后的热力图,可以看到添加混合注意力机制后算法着重关注待检测目标,抗干扰能力较强,证明了混合注意力机制的有效性。

3.5　自制吸烟数据集实验结果分析

自制吸烟数据集中包含了各种复杂情况下的吸烟目标,本文通过对比各类算法在自制吸烟数据集上的表现,测试 Tr-SSD 算法的泛化能力,实验数据对比结果如表 5 所示。由表 5 可知,Tr-SSD 算法的 *mAP* 值在自制吸烟数据集中取得了最优水平,因此,Tr-SSD 算法具有更好的鲁棒性与泛化性。



(a) 未添加混合注意力机制 (b) 添加混合注意力机制

图7 RSOD数据集热力图对比

Figure 7 Heat map comparison of RSOD data sets

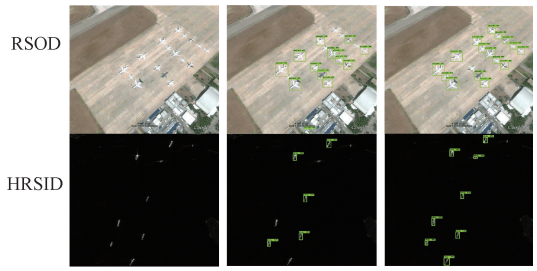
表5 自制吸烟数据集定量实验对比

Table 5 Quantitative experimental comparison of homemade smoking data sets

算法	输入尺寸/像素	骨干网络	*mAP*/%
Faster RCNN	1 000×600	VGG16	75.1
SSD	300×300	VGG16	78.5
RSSD	300×300	VGG16	79.9
DSSD	300×300	Resnet-101	80.2
IPG-Net	1 000×600	IPGnet-101	84.9
Tr-SSD	300×300	Resnet50	86.7

3.6 定性实验结果分析

为了更直观的表示本文算法在实际应用中的检测效果,本文选取RSOD数据集与HRSID数据集进行算法的定性实验,如图8所示。



(a) 待检测图片 (b) SSD检测效果 (c) Tr-SSD检测效果

图8 定性实验效果对比

Figure 8 Comparison of qualitative experimental effects

图8(a)中RSOD图片包含17个待检测目标,HRSID图片包含8个待检测目标。图8(b)为SSD算法的检测效果,在RSOD图片中,SSD算法共检测出13个目标,漏检5个目标,错检1个目标;在HRSID图片中,SSD算法共测出5个目标,漏检3个目标。这是因为SSD算法在提取小目标图像特征的过程中,易受到背景中的杂乱信息干扰,出现错检与漏检。由图8(c)可知,Tr-SSD算法在2张待检测图片中分别检测出了16个、8个目标,且没有出现错检目标。

Tr-SSD算法对于密集小目标的检测精度明显优于SSD算法,且Tr-SSD算法的错检率、漏检率也明显低于SSD算法,这是由于Tr-SSD算法使用了基于Transformer的特征融合策略与混合注意力机制,增强了小目标的位置信息,并且抑制了杂乱背景下的无用信息。

4 结论

本文针对小目标检测任务的检测难度较大、检测精度较低等问题,提出了一种结合混合注意力机制与多尺度特征融合的小目标检测的Tr-SSD算法。算法通过替换原本的特征提取网络来减少特征提取时的信息丢失,提高特征提取能力,并且在网络的中尺度特征图中通过设计混合注意力增强有效信息的表达,同时建立不同尺度特征图中的远距离依赖。在中高层特征图中设计了一个特征融合模块,将相邻特征图进行融合,以增强特征图的语义信息。实验结果表明,Tr-SSD算法精度优于当前目标检测领域性能较好的算法,同时能够满足实时检测要求,可用于实时目标检测任务。下一步将进行网络的轻量化,进一步提高网络的检测速度。

参考文献:

[1] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot MultiBox detector[J]. Computer Vision, 2016, 9905: 21-37.

[2] FU C Y, LIU W, RANGA A, et al. DSSD: deconvolutional single shot detector[EB/OL]. (2017-01-23)[2023-09-17]. https://arxiv.org/abs/1701.06659.

[3] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2017: 936-944.

[4] TAN M X, PANG R M, LE Q V. EfficientDet: scalable and efficient object detection[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 10778-10787.

[5] QIAO S Y, CHEN L C, YUILLE A. DetectoRS: detecting objects with recursive feature pyramid and switchable atrous convolution[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021: 10208-10219.

[6] LIM J S, ASTRID M, YOON H J, et al. Small object detection using context and attention[C]//2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC). Piscataway: IEEE, 2021: 181-186.

[7] YANG C, HUANG Z H, WANG N Y. QueryDet: cascaded sparse query for accelerating high-resolution small object detection[C]//2022 IEEE/CVF Conference on Com-

puter Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 13658-13667.
[8] REDMON J, FARHADI A. YOLOv3: an incremental improvement[EB/OL]. (2018-04-08)[2023-09-17]. https://arxiv.org/abs/1804.02767.
[9] 马学森, 马吉, 蒋功辉, 等. 基于注意力机制和多尺度特征融合的绝缘子缺陷检测方法[J]. 南京大学学报(自然科学), 2022, 58(6): 1020-1029.
MA X S, MA J, JIANG G H, et al. Insulator defect detection method based on attention mechanism and multi-scale feature fusion[J]. Journal of Nanjing University (Natural Science), 2022, 58(6): 1020-1029.
[10] QU Z, HAN T Q, YI T M. MFFAMM: a small object detection with multi-scale feature fusion and attention mechanism module[J]. Applied Sciences, 2022, 12(18): 8940.
[11] NI J, WANG R, TANG J. ADSSD: improved single-shot detector with attention mechanism and dilated convolution[J]. Applied Sciences, 2023, 13(6): 4038.
[12] SHAN D R, XU Y L, ZHANG P, et al. DPSSD: dual-path single-shot detector[J]. Sensors, 2022, 22(12): 4616.
[13] SHI H, CHAI B Q, WANG Y P, et al. A local-sparse-information-aggregation transformer with explicit contour guidance for SAR ship detection[J]. Remote Sensing, 2022, 14(20): 5247.
[14] ZHANG X S, WAN F, LIU C, et al. Learning to match anchors for visual object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(6): 3096-3109.
[15] SHI H, FANG Z H, WANG Y P, et al. An adaptive sample assignment strategy based on feature enhancement for ship detection in SAR images[J]. Remote Sensing, 2022, 14(9): 2238.
[16] YU J M, WU T, ZHANG X, et al. An efficient lightweight SAR ship target detection network with improved regression loss function and enhanced feature information expression[J]. Sensors, 2022, 22(9): 3447.
[17] LIU M J, WANG X H, ZHOU A J, et al. UAV-YOLO: small object detection on unmanned aerial vehicle perspective[J]. Sensors, 2020, 20(8): 2238.
[18] HUANG Z C, WANG J L, FU X S, et al. DC-SPP-YOLO: dense connection and spatial pyramid pooling based YOLO for object detection[J]. Information Sciences, 2020, 522: 241-258.

Small Object Detection Based on Feature Fusion and Mixed Attention

WEI Mingjun[1,2], WANG Mohan[1], LIU Yazhi[1,2], LI Hui[1]

(1. College of Artificial Intelligence, North China University of Science and Technology, Tangshan 063210, China; 2. Hebei Provincial Key Laboratory of Industrial Intelligent Perception, North China University of Science and Technology, Tangshan 063210, China)

Abstract: To address to the low feature information, low detection rates, and high false rate and missing rate in the target detection task, a Tr-SSD algorithm based on multiscale feature fusion and a hybrid attention mechanism was proposed. Firstly, a Resnet50 residual network was utilized as the backbone network for the SSD algorithm to enhance its feature extraction capabilities. Secondly, a hybrid attention mechanism was designed and applied to the mid-scale feature maps of the network to enhance effective information within the feature maps and establish long-range dependencies between pieces of information. Finally, a FPN (feature pyramid network) structure was formed by using network layers centered around the Transformer instead of the original backbone network in the SSD algorithm, which fused feature information of different scales to more accurately locate small targets. Experimental results showed that the Tr-SSD algorithm achieved *mAP* values of 81.9%, 87.5%, and 88.4% on the PASCAL VOC dataset, HRSID dataset, and RSOD remote sensing dataset, respectively. This represented an improvement of 4.7 percentage points, 6.8 percentage points, and 9.2 percentage points compared to the original SSD algorithm. Moreover, the detection speed could meet the requirements for real-time detection.

Keywords: small target detection; attention mechanism; feature fusion; deep learning; real-time detection