控制与决策
Control and Decision

全局与局部图像特征自适应融合的小目标检测算法

赵亮, 刘世鹏

引用本文:
赵亮,刘世鹏. 全局与局部图像特征自适应融合的小目标检测算法[J]. 控制与决策, 2023, 38(4): 935–943.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2021.1800

全局与局部图像特征自适应融合的小目标检测算法

赵 亮[1,2†], 刘世鹏[1]

(1. 西安建筑科技大学 信息与控制工程学院，西安 710055；
2. 陕西省岩土与地下空间工程重点实验室，西安 710055)

摘 要: 针对现有目标检测算法对于小目标检测精度低的问题,提出一种全局与局部图像特征自适应融合的一阶段小目标检测算法SODet. 首先,将Transformer与卷积神经网络相结合构建主干网络,分别提取图像全局和局部信息,并利用自适应特征选择模块AFS对二者输出进行融合;然后,在特征融合网络中利用额外尺度特征图进行特征融合,同时利用大目标抑制单元约束大目标特征表达、转移小目标特征,输出4个尺度的特征图送入预测网络;最后,在损失函数部分针对小目标检测利用EIOU和Focal loss进行优化. 实验结果表明,SODet算法在MS COCO验证集上AP_S达到31.5%,相比于其他算法具有较强的竞争力,同时具有较高的推理速度.

关键词: 小目标检测；Transformer；EIOU；Focal loss；FPN

中图分类号: TP183 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2021.1800

引用格式: 赵亮,刘世鹏. 全局与局部图像特征自适应融合的小目标检测算法[J]. 控制与决策, 2023, 38(4): 935-943.

Small object detection algorithm based on adaptive fusion of global and local image features

ZHAO Liang[1,2†], *LIU Shi-peng*[1]

(1. College of Information and Control Engineering，Xi'an University of Architecture and Technology，Xi'an 710055，China；2. Shaanxi Provincial Key Laboratory of Geotechnical and Underground Space Engineering，Xi'an 710055，China)

Abstract: Aiming at the problem that the existing object detectors have low accuracy for small objects. A one-stage small object detector (SODet) is proposed to adaptively fuse global and local image features. Firstly, the Transformer and the convolutional neural network (CNN) are combined to construct a backbone network to extract global and local information of the image respectively. Then the adaptive feature selection (AFS) module is used to fuse the outputs of the Transformer and CNN. Then, extra-scale feature maps are adopted in the feature fusion network. At the same time, the large object restraint unit is applied to constrain the expression of large object features and transfer small object features. The feature maps of four scales are sent to the prediction network. Finally, in the loss function, the EIOU and Focal loss are used to optimize small object detection. The experimental results show that the SODet has 31.5% in terms of AP_S on the MS COCO verification set, which is more competitive than other algorithms and has a higher inference speed.

Keywords: small object detection；Transformer；EIOU；Focal loss；FPN

0 引 言

目标检测[1-2]是计算机视觉领域中的基本问题之一,其中比较困难的小目标检测[3-5]更是近些年研究的热点. 小目标检测的难度高是因为该类目标在图像中像素占比低,且常与背景或其他目标混杂在一起形成难以辨识的特征,从而造成检测困难. 此外,任何尺度的目标在卷积神经网络(CNN)中由浅层传递至深层的过程中均会丢失部分细节和边缘信息,特别是小目标在传递至网络深层时其信息甚至消失,故而小目标检测精度远远落后于其他尺度目标[6-9]. 随着CNN的发展,其在网络深度[10]、广度[11]等方面出现了具有较强特征提取能力的网络结构变体以及能够极大增加其感受野的膨胀卷积[12]和金字塔池化[13]操作,但CNN仍然受限于卷积核的局

收稿日期: 2021-10-20；**录用日期:** 2022-03-15.

基金项目: 国家自然科学基金项目(51209167, 12002251)；陕西省自然科学基金项目(2019JM-474)；西安市科技计划项目(2020KJRC0055)；陕西省岩土与地下空间工程重点实验室开放基金项目(YT202004).

责任编委: 柴利.

†通讯作者. E-mail: zhaoliang@xauat.edu.cn.

部计算方式, 导致无法获取图像中远程目标之间的依赖关系和全局信息. 为了克服上述问题, 现有小目标检测算法多通过添加注意力机制[14-17] 以提升CNN的特征提取能力, 但想要捕获远程目标之间的空间关系和全局信息依然是本领域难点. 最近, 有学者利用Transformer[18] 模型对整张图像进行计算以捕获图像的全局信息, 在图像分类和目标检测等任务性能表现良好. Dosovitskiy等[19] 提出了一种视觉Transformer模型(ViT), 将输入图像切分为固定大小的图像块作为words送入Transformer的编码器并使用全连接层进行类别预测, 结果表明ViT在图像分类任务上表现良好. Liu等[20] 提出了具有层次化结构的Swin Transformer作为检测算法主干网络, 其能够像CNN一样逐渐降低图像分辨率并增大感受野, 再加上移位窗口操作, Swin-B模型在ImageNet-1 K上获得了83.3 %的Top-1精度, 性能几乎媲美EfficientNet系列, 同时参数量最大仅有88 M. Zhang等[21] 提出了一种具有高效Transformer模块、结构简单的层次化Transformer模型ResT, 在ImageNet同样取得具有竞争力的结果. 虽然具有全局计算特性的Transformer模型总体性能较强, 但存在丢失部分局部信息、对小目标不敏感的问题, 故而对小目标的检测效果较差.

针对上述问题, 本文提出一种基于Transformer和CNN的一阶段目标检测算法SODet. 具体内容有: 1) 在主干网络利用CNN和Transformer分别提取图像局部和全局信息, 并设计自适应特征选择模块对二者输出进行自适应融合. 2) 针对小目标检测, 在特征融合网络利用额外尺度特征图进行特征融合, 同时设计大目标抑制单元对大目标特征进行约束、转移小目标特征, 将4个尺度的输出特征图送入预测网络进行预测. 3) 在损失函数部分采用EIOU[22] 和Focal loss[9] 优化小目标检测.

1 SODet目标检测算法

1.1 SODet算法整体结构

SODet算法由主干网络CoHiT (convolutional neural networks and hierarchical Transformers)、特征融合网络FrtPN (feature restrict and transfer pyramid network) 和预测网络3部分组成, 图1为SODet算法的整体结构示意图. CoHiT利用具有层次化结构的Transformer (Trans) 和CNN (CNNs) 捕获图像全局和局部特征, 并利用自适应特征选择模块AFS (adaptive feature selection) 对二者输出进行有效融合. FrtPN能够抑制大物体特征和转移小物体特征, 同时融合更多尺度的主干网络输出特征. 预测网络采用与YOLOv4[23] 相同的检测头, 在4个尺度上进行回归和分类任务.

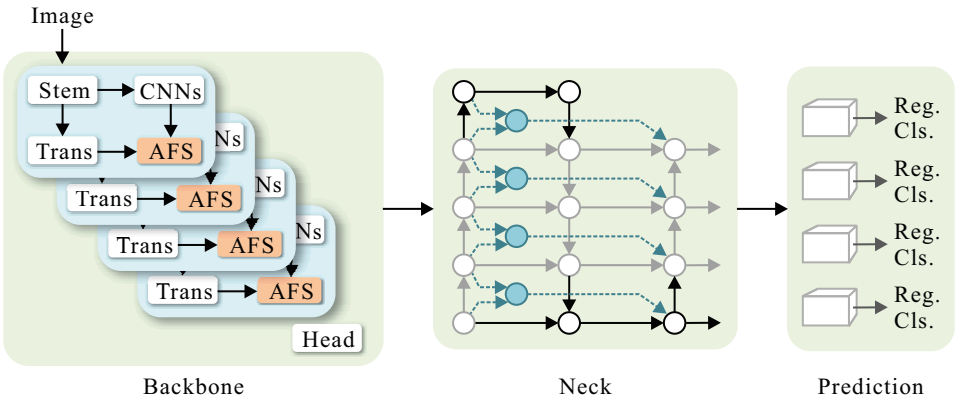


图 1 SODet算法整体结构

1.2 主干网络CoHiT

在图像中距离相对较远的目标间可能依然存在上下文关系, 但是具有有限且固定大小感受野的CNN在捕获上述关系时难以得到有效的结果, 而具有全局计算特性的Transformer能够在一定程度上满足该方面的需要. 据此本文提出一种将CNN与具有层次化Tansformer结构的ResT模型[21] 进行结合的主干网络CoHiT, 同时捕获图像中的局部信息和全局信息.

图2为CoHiT网络结构示意图, 由Stem模块、阶段化特征提取模块(Stage 1 ~ Stage 4) 和Head模块3部分组成, 输入图像经由Stem模块进行预处理后送入阶段化特征提取模块提取图像特征, 之后在网络深层利用卷积结构的Head模块得到高级语义特征. Stem模块采用1×1卷积和步长为2的3×3卷积对输入图像进行预处理, 进行单次下采样同时保留更多细节和纹理信息. 阶段化特征提取模块中每个阶段由CNN模块和Transformer模块提取图像特征, 利

用AFS将两个分支的输出特征进行自适应融合并送入下一阶段,其中Transformer的输出结果需经过层规范化(Layer Normalization)再送入AFS,CNN模块包含两个1×1卷积层、一个3×3卷积层、批规范化(Batch Normalization)和LeakyReLU激活函数.

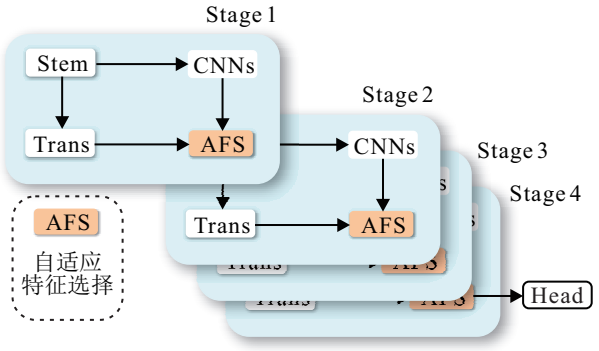


图 2　CoHiT网络结构

由于CNN和Transformer各自不同的计算特性,采用直接相加的方式无法有效地融合局部信息与全局信息,进而影响后续阶段的特征提取. 针对此问题,本文提出一种自适应特征选择模块AFS用于特征融合,该模块利用注意力机制设置空间门控单元G_S和通道门控单元G_C,分别用于选择空间信息和通道信息,如图3(a)所示,AFS结构分为上下空间门控路径和中间通道门控路径,在当前阶段(Stage S)中CNN模块和Transformer模块的输出X_C和X_T分别作为上下两路径的输入,利用空间门控单元G_S计算得到中间特征U和V,即

$$U/V = G_S(X_C/X_T), \tag{1}$$

其中X_C、X_T、U和V为三维数组且$X_C, X_T, U, V \in \boldsymbol{R}^{H\times W\times C}$. 中间路径对$X_C$和$X_T$进行逐点相加得到综合特征$X$,再利用通道门控单元$G_C$计算得到用于选择通道信息的决策向量$\lambda(1-\delta)$和$\lambda(\delta)$,有

$$\lambda(1-\delta) = (1-\delta)(G_C(X_C + X_T)), \tag{2}$$

$$\lambda(\delta) = \delta(G_C(X_C + X_T)), \tag{3}$$

其中δ为Softmax函数. 将决策向量$\lambda(1-\delta)$和$\lambda(\delta)$分别与中间特征U和V逐点相乘再相加得到最终输出特征图Y,之后送入在下一阶段(Stage $S+1$)的CNN模块和Transformer模块用于后续特征提取. 当前阶段Y的计算过程如下式所示:

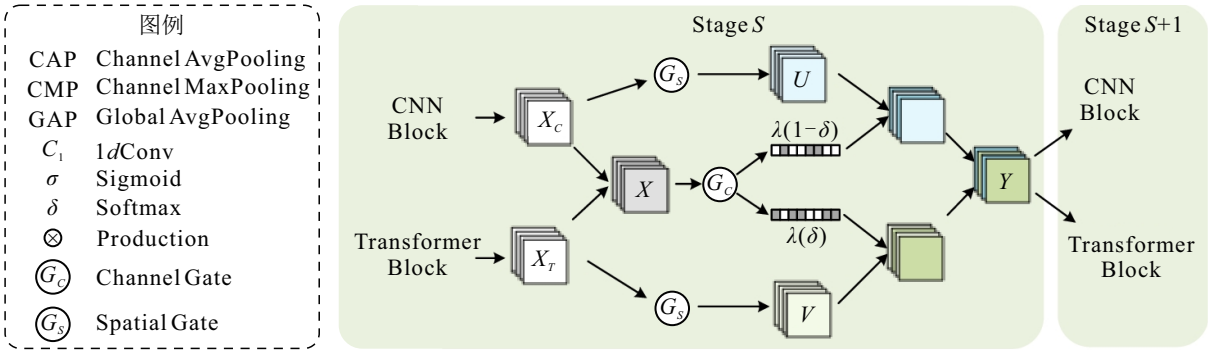
$$Y = U\cdot\lambda(1-\delta) + V\cdot\lambda(\delta), \tag{4}$$

其中Y与X_C/X_T尺寸相同,即$Y \in \boldsymbol{R}^{H\times W\times C}$. AFS上下路径的空间门控单元结构如图3(b)所示,沿通道维度对X_C/X_T作均值池化和最大池化得到两个$H\times W$矩阵,送入卷积层进行特征提取得到二维权重矩阵,由Sigmoid函数归一化后再与X_C/X_T进行矩阵乘积即可得到中间特征U或V,该过程如下式所示:

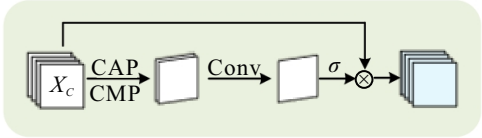
$$U/V = X_C/X_T\cdot\sigma(C_{7\times7}(P_a^C(X_C/X_T)), P_m^C(X_C/X_T)). \tag{5}$$

其中: P_a^C和P_m^C分别为通道维度的均值池化和最大池化,$C_{7\times7}$表示卷积核大小为7×7的卷积层. 在中间路径的通道门控单元中(如图3(c)所示),对综合特征X作空间维度的全局最大池化得到尺寸为$1\times1\times C$的一维向量,为了减少计算量,采用一维卷积对向量进行特征提取得到权重向量v,即

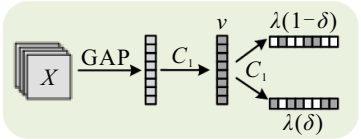
$$v = C_5^{1d}(P_m^S(X)). \tag{6}$$



(a)　自适应特征选择模块AFS



(b)　空间门控单元



(c)　通道门控单元

图 3　自适应特征选择模块

其中: C_5^{1d} 表示一维卷积, 卷积核大小为5; $v \in \boldsymbol{R}^{1\times1\times C}$; P_m^S 为空间维度全局最大池化算子. 再利用两个全连接层 fc 对 v 计算得到两个向量 x 和 y, 全连接层参数分别为 W^x 和 W^y, 有

$$x, y = v \cdot fc[W^x, W^y]. \tag{7}$$

之后利用Softmax函数对 x 和 y 进行归一化, 使得两个向量中相同位置的元素 x^C、y^C 之和为1, 有

$$\begin{cases} y^C = \delta(y^C, x^C), \\ x^C = 1 - y^C. \end{cases} \tag{8}$$

其中 x^C 和 y^C 分别为向量 x 和 y 的第 C 个元素, 其总数等于 X 的通道数 C. 将所有元素进行拼接得到决策向量 $\lambda(1-\delta)$ 和 $\lambda(\delta)$, 即

$$\begin{cases} \lambda(1-\delta) = [x^1, x^2, \ldots, x^C], \\ \lambda(\delta) = [y^1, y^2, \ldots, y^C]. \end{cases} \tag{9}$$

Head模块由4组CNN模块构成, 能够将来自阶段化特征提取模块的输出特征进行有效融合同时提取高级语义特征. CoHiT网络结构配置如表1所示, 固定每个阶段中Transformer模块(Trans, T)数量, 利用不同数量的CNN模块(CNNs, C)构建不同复杂度的网络模型, 按照模型复杂度由小至大依次为 S、M、L 和 X 四种主干网络.

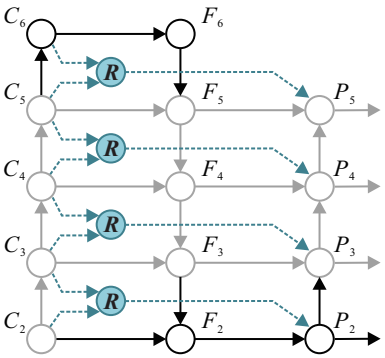
表 1 CoHiT网络结构配置

Stage	S	M	L	X
1	[Trans]×2 [CNNs]×1	$[T]\times2$ $[C]\times2$	$[T]\times2$ $[C]\times3$	$[T]\times2$ $[C]\times3$
2	[Trans]×2 [CNNs]×2	$[T]\times2$ $[C]\times3$	$[T]\times2$ $[C]\times4$	$[T]\times2$ $[C]\times4$
3	[Trans]×6 [CNNs]×3	$[T]\times6$ $[C]\times4$	$[T]\times6$ $[C]\times5$	$[T]\times6$ $[C]\times23$
4	[Trans]×2 [CNNs]×1	$[T]\times2$ $[C]\times2$	$[T]\times2$ $[C]\times3$	$[T]\times2$ $[C]\times3$

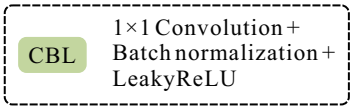
1.3 特征融合网络FrtPN

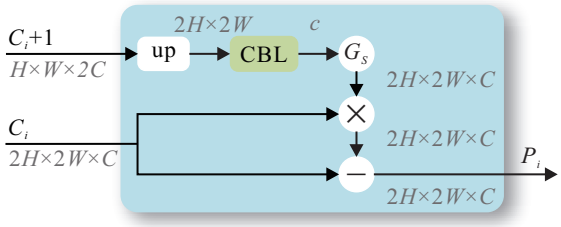
在目标检测领域中, 常见的特征金字塔网络[24-25]虽然能够通过融合深层特征与浅层特征, 采用不同尺度的特征图进行预测获得部分收益, 但是小目标检测的性能提升相对较少, 为此, 本文提出一种能够对小目标特征作单独处理的特征金字塔网络结构FrtPN, 以此增强算法对小目标的检测效果, 图4(a)为FrtPN网络结构示意图.

FrtPN网络在PAFPN[24]的基础上分别在更深层和更浅层添加额外尺度特征图进行融合, 即采用5个不同尺寸大小的特征图进行多尺度融合. 图4(a)中,



(a) FrtPN整体结构





(b) 大目标抑制单元

图 4 FrtPN网络结构

深层的 C_5 为主干网络Head模块输出特征图, FrtPN网络再次利用与Head相同结构对 C_5 进行处理, 将其空间分辨率降低1倍、通道数不变; 之后通过卷积层处理得到 F_6, 并将其上采样与 F_5 进行融合, 该过程中无需通道降维; 浅层方面, C_2 为主干网络Stage 2的输出特征图, 经处理后得到 F_2、P_2 并与其他节点相连.

图像信息在网络中传递时, 深层特征图包含的小目标信息少于浅层特征图. 根据该特性, 在 $C_2 \sim C_6$ 的每相邻2个特征图之间嵌入大目标抑制单元(large objects restraint unit, LRU), 即图4(a)中的 R, LRU利用网络深层特征图中的大目标信息抑制浅层中大目标的特征表达, 从而保留浅层特征图中小目标的特征, 并将输出连接至预测层 $P_2 \sim P_5$, 图4(b)为其实现细节. LRU包含2个输入端 C_{i+1}、C_i 和1个输出端 P_i, 其中 C_{i+1} 为深层特征图且 $C_{i+1} \in \boldsymbol{R}^{H\times W\times 2C}$, C_i 和 P_i 为较浅一层的特征图且 $C_i, P_i \in \boldsymbol{R}^{2H\times 2W\times C}$. C_{i+1} 输入LRU后首先利用最近邻插值法上采样将其空间分辨率变为 $2H \times 2W$, 再利用CBL模块对其通道从 $2C$ 降维至 C (C_6 不进行降维处理), 其中CBL包含1×1卷积层、批规范化和LeakyReLU激活函数. 之后利用空间门控单元 G_S 进行后处理输出维度为 $2H \times 2W \times C$ 的特征图 T_R, 即

$$T_R = G_S(\mathrm{CBL}(\mathrm{up}(C_{i+1}))), \tag{10}$$

其中up为上采样操作. 之后利用Sigmoid函数进行归一化,得到大目标的特征尺度函数$S(T_R)$,且$S(T_R)\in(0,1)$,有

$$S(T_R)=\frac{1}{1+\mathrm{e}^{-T_R}}. \tag{11}$$

将尺度函数与C_i逐点相乘即可增强原本的大目标特征表达,再与C_i作差即可消除部分大目标特征,从而保留小目标特征,最后将输出结果以逐点相加方式嵌入P_i得到新的特征图P_i^*,如下式所示:

$$P_i^*=(C_i-C_i\cdot S(T_R))+P_i. \tag{12}$$

1.4 损失函数

在损失函数设计阶段,一些目标检测算法在进行预测框回归时,往往采用CIOU[23]损失函数,通过计算预测框回归误差,分别考虑预测框B和真实框B^{gt}的重叠面积、中心点距离和宽高比得到预测框回归误差,宽高比仅能反映预测框与真实框模糊差异,而非它们的实际误差. 本文在SODet算法损失函数部分引入EIOU[22]函数以解决该问题,EIOU函数表达式为

$$L_{\mathrm{EIOU}}=L_{\mathrm{IOU}}+\frac{\rho^2(B_{ctr},B_{ctr}^{gt})}{c^2}+\frac{\rho^2(w,w^{gt})}{c_w^2}+\frac{\rho^2(h,h^{gt})}{c_h^2}. \tag{13}$$

其中: $L_{\mathrm{IOU}}=1-\dfrac{|B\bigcap B^{gt}|}{|B\bigcup B^{gt}|}$, $\rho(\cdot)$为欧氏距离, B_{ctr}^{gt}、w^{gt}和h^{gt}分别为真实框的中心点坐标、宽和高, B_{ctr}、w和h分别为预测框的中心点坐标、宽和高, c_w和c_h分别为预测框和真实框的最小外接矩形的宽度和高度. 相较于CIOU函数利用宽高比反映包围框的模糊差异,EIOU函数利用$\rho^2(w,w^{gt})/c_w^2$和$\rho^2(h,h^{gt})/c_h^2$直接计算包围框宽高的实际误差同时能够进一步加快收敛速度. SODet算法的损失函数由回归误差L_{EIOU}、置信度误差L^{obj}、L^{noobj}和预测类别误差L^{cls}三部分组成,如下式所示:

$$L=L_{\mathrm{EIOU}}+L^{\mathrm{obj}}+L^{\mathrm{noobj}}+L^{\mathrm{cls}}. \tag{14}$$

此外,相比于较大尺度的目标,算法在检测图像中的小目标时往往比较困难,本文利用Focal loss对式(14)进行改进,以降低较易分类目标的误差比重、增加较难分类的误差比重,从而提升算法对小目标的检测性能,如下式所示将$L^*:(L^{\mathrm{obj}},L^{\mathrm{noobj}},L^{\mathrm{cls}})$修改为$L_{\mathrm{focal}}^*$:

$$L_{\mathrm{focal}}^*=(1+\varepsilon-e^{-L^*})^{\gamma}\cdot L^*. \tag{15}$$

其中: $\gamma=2,\varepsilon$为一个极小的正数用以防止误差值为0. 最终SODet 算法的损失函数如下式所示:

$$L=L_{\mathrm{EIOU}}+L_{\mathrm{focal}}^{\mathrm{obj}}+L_{\mathrm{focal}}^{\mathrm{noobj}}+L_{\mathrm{focal}}^{\mathrm{cls}}. \tag{16}$$

2 实验验证和分析

2.1 消融实验和分析

实验平台为Ubuntu 16.04,实验数据集为MS COCO,评价指标采用平均精度AP和小目标平均精度$\mathrm{AP_S}$,输入图像大小为608×608,所有模型均在ImageNet进行预训练,在MS COCO进行微调,采用AdamW[26]优化算法更新参数,训练周期为96 epochs,设置初始学习率为0.001,在第64 epoch和第88 epoch分别衰减10倍.

2.1.1 主干网络CoHiT消融实验和分析

SODet算法主干网络利用CNN和Transformer提取图像特征,并利用自适应特征选择模块AFS融合二者输出,为了验证CoHiT的有效性,本文以SODet为基线模型,利用ResNet-50[10](CNN)、 ResT-B[21](Trans)和CoHiT-S (Hybrid, CNN+Trans)作为主干网络进行实验验证.

实验结果如表2所示,CoHiT-S作为主干网络,以直接相加(Add.)的方式融合CNNs与Transformers的输出时,AP和$\mathrm{AP_S}$均比ResNet-50和ResT-B高,但相较于ResT-B小目标检测精度$\mathrm{AP_S}$的增益仅有0.9%;而采用AFS融合方式的CoHiT-S比直接相加的AP和$\mathrm{AP_S}$分别提升了1.6%和0.7%,$\mathrm{AP_S}$比ResT高1.6%. 实验结果表明,CoHiT能够有效地融合CNNs与Transformers的输出结果,且有利于小目标检测.

表 2　主干网络CoHiT消融实验结果

method	Backbone	Fusion	AP/%	$\mathrm{AP_S}$/%
CNN	ResNet-50	–	33.4	20.1
Trans	ResT-B	–	38.0	22.5
Hybrid	CoHiT-S	Add.	39.2	22.5
Hybrid	CoHiT-S	AFS	40.8	24.1

CoHiT网络的小目标检测精度相较于单独CNN或Transformer网络算法有所提升. 本文以Mask R-CNN[27]为基线算法与其他CNN融合Transformer的主干网络进行对比实验,训练采用MS COCO数据集,设置批次大小为32,优化器采用AdamW,训练周期为12 epochs,初始学习率为0.000 1,在第8 epoch和第11 epoch时学习率衰减10倍. 实验结果如表3所示,结果表明,CoHiT主干网络对于小目标的检测性能较好,尤其是CoHiT-X的$\mathrm{AP_S}$和AP分别达到了29.6%和47.6%,远高于Conformer[28]与DS-Net[29],且具有较低网络复杂度的CoHiT网络同样具有较强竞争力.

表 3 不同融合主干网络性能对比

Backbone	input	params / M	AP / %	AP_S / %
ResNet-101	1 333×800	63.2	40.0	22.6
Conformer[28]	1 120×800	56.9	44.9	28.7
DS-Net[29]	1 333×800	43.2	44.3	28.3
CoHiT-S	1 333×800	44.2	43.2	27.6
CoHiT-M	1 333×800	51.1	44.8	28.3
CoHiT-L	1 333×800	58.1	46.1	28.8
CoHiT-X	1 333×800	77.2	47.6	29.6

2.1.2 特征融合网络 FrtPN 消融实验和分析

大目标抑制模块 LRU 能够利用深层特征图对浅层特征图中的大目标特征进行抑制, 从而传递并增强浅层中的小目标特征, 本文将 C_2 与 C_3 间的 LRU 内部运行过程进行可视化, 如图 5 所示. 主干网络对原始图像进行特征提取得到特征图 C_2 (150 × 150 × 256) 和 C_3 (75 × 75 × 512), C_2 比 C_3 包含更多小目标特征 (棒球部分). C_3 经过上采样、CBL 模块和空间门控单元 G_S 处理后得到与 C_2 维度相同的特征图, 同时其特征表达也更加集中, 之后经过 Softmax 函数归一化后作为缩放函数与 C_2 进行相乘, 能够提升 C_2 大目标特征 (人体部分) 同时削减小目标特征, 最后经过与原特征图 C_2 作差即可对原本大目标特征进行抑制, 同时保留较为丰富的小目标特征.

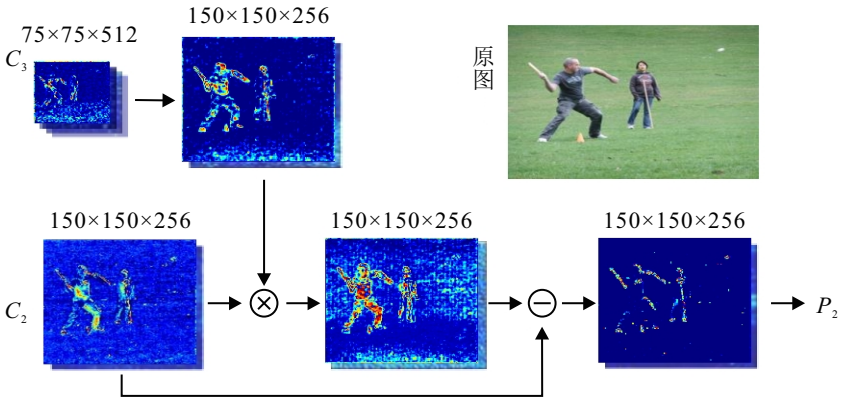


图 5 LRU 模块内部运行过程

为了验证 FrtPN 对小目标检测的有效性, 以 SODet 和 CoHiT-S 为基线模型进行消融实验, 采用 C_6 和 C_2 为对应的额外尺度特征图, P_2 为采用 4 个尺度进行预测, LRU 为大目标抑制单元, √表示对应方法被使用, M.x 为对应方法名称.

实验结果如表 4 所示, PAFPN (M.0) 的 AP 和 AP_S 分别为 37.5 % 和 24.7 %, 添加 C_6 (M.1) 或 C_2 (M.2) 额外尺度用于特征融合后精度均有所提升, 但采用 C_2 带来的增益更多, 表明浅层网络特征更加有利于小目标检测, 同时添加 C_6 和 C_2 时 (M.3) 小目标检测精度 AP_S 达到了 25.4 %; 在 M.1、M.2 和 M.3 的基础上添加 LRU 模块 (M.1†、M.2† 和 M.3†), AP_S 取得进一步提升, 分别达到 25.4 %、26.0 % 和 26.2 %; M.4 在 M.3 基础上添加了更浅层尺度的特征图用于预测, AP_S 取得了 1.1 % 的大幅提升, 表明在更大特征图进行预测有利于小目标检测, 此外在其基础上添加 LRU 模块 (M.4†) 后, AP 和 AP_S 均达到了最高即 42.0 % 和 27.3 %, 充分表明大目标抑制模块能够约束大目标特征、增强小目标的特征表达.

表 4 特征融合网络 FrtPN 消融实验结果

name	C_6	C_2	P_2	LRU	AP / %	AP_S / %
M.0					37.5	24.7
M.1	√				37.8	24.8
M.1†	√			√	38.6	25.4
M.2		√			38.1	25.3
M.2†		√		√	39.1	26.0
M.3	√	√			38.9	25.4
M.3†	√	√		√	40.1	26.2
M.4	√	√	√		40.8	26.5
M.4†	√	√	√	√	42.0	27.3

2.1.3 损失函数消融实验和分析

本文在 SODet 算法损失函数中引入 EIOU 函数用于预测框回归, 同时针对小目标检测采用 Focal loss 优化分类损失, 降低较易分类的误差比重, 增高较难分类的误差比重. 为了验证损失函数的效果, 以 SODet 和 CoHiT-S 为基线模型, 采用多种损失函数的组合进行消融实验, 采用 Cls. 和 Reg. 为分类和回归损失函数, CE 为交叉熵函数, MSE 为均方误差函数, FL 为 Focal loss 函数, M.x 为对应方法名称.

实验结果如表 5 所示, M.a 为 CE 与 MSE 的组合, 其 AP 和 AP_S 分别为 35.2 % 和 20.5 %, 采用 CIOU 替换 MSE 作为回归部分时, 精度均有所提升; 在 M.b 中引入 EIOU 后 AP 和 AP_S 分别达到了 38.0 % 和 22.7 %, 表明精细化的计算包围框之间的差异是有效的; 将分类损失函数部分利用 Focal loss 优化后 (M.d), AP

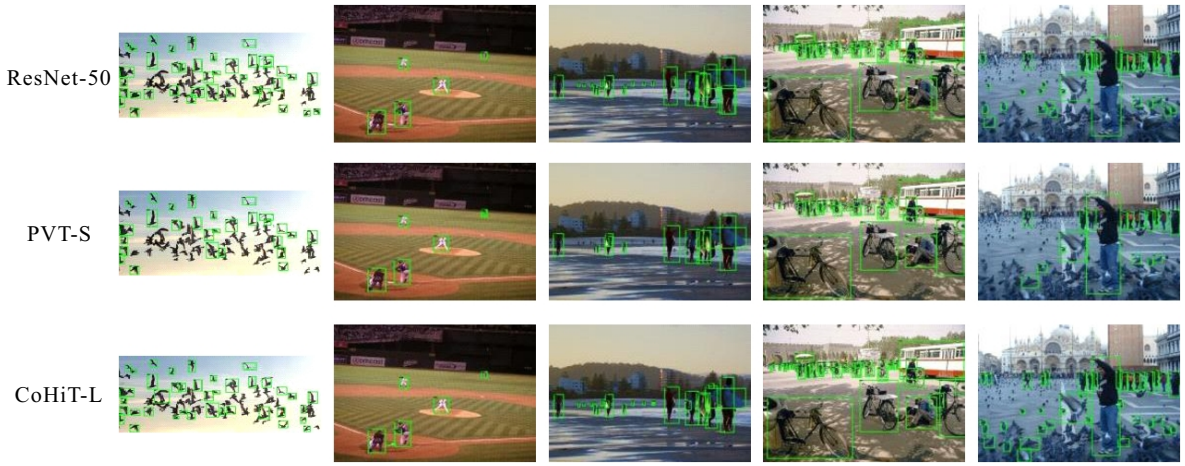
和 AP_S 分别得到了 1.8% 和 0.9% 的大幅提升, 表明 Focal loss 能够有效提升较难分类目标的训练效果.

表 5　损失函数消融实验结果

name	Cls.	Reg.	AP/%	AP_S/%
M.a	CE	MSE	35.2	20.5
M.b	CE	CIOU	37.5	21.4
M.c	CE	EIOU	38.0	22.7
M.d	FL	EIOU	39.8	23.6

2.2　SODet算法对比实验和检测结果分析

为了直观地展示 SODet 算法的检测结果, 本文选取 MS COCO 测试集的图片进行检测并将结果进行可视化, 如图 6 所示. 以 SODet 算法为基线框架, 分别将 ResNet-50、PVT-S[30] 和 CoHiT-L 作为主干网络进行检测. 相较于 ResNet-50 与 CoHiT-L, PVT-S 的检测结果中丢失了部分小目标 (图 6(a) 中相对较小的鸟、图 6(c) 中远处的人), 表明 Transformer 结构对小目标的敏感性较低, 进而造成小目标的漏检现象. 此外, CoHiT-L 的结果中得到的包围框数量更多 (图 6(a) 中相对较小的鸟和图 6(e) 中靠近建筑物微小的人), 表明 CNN 与 Transformer 结合能够提升对小目标的检测效果.



(a) 第 1 组　(b) 第 2 组　(c) 第 3 组　(d) 第 4 组　(e) 第 5 组

图 6　MS COCO 检测结果可视化

为了进一步验证 SODet 算法对于小目标的检测性能, 在 MS COCO 数据集与当下其他主流算法进行对比. 实验设备采用 4 张 NVIDIA Geforce 3 090 GPU 并行训练, 1 张 3 090 GPU 进行推理测试, 训练样本利用 Mosaic 和 Mixup 进行数据增强, 输入图像大小设置为 608×608, 训练周期为 200 epochs, 采用 AdamW 优化算法更新网络参数, 权重衰减指数为 0.05, 批次大小为 16, 设置初始学习率为 0.000 1, 采用余弦退火策略更新学习率.

实验结果如表 6 所示, 其中 MS^{train} 为多尺度训练策略. 以 CoHiT-S 为主干网络的 SODet 算法在 608×608 的输入尺寸时, AP_S 达到了与具有更大输入尺寸、更高网络复杂度的 SAPD[31] 相当的水平, 同时

表 6　MS COCO 验证集算法性能对比

method	Backbone	image size	MS^{train}	FPS	AP/%	AP_S/%	year
ATSS[37]	R 101	1 333×800	√	15	43.6	26.1	2020
YOLOv4[23]	CD 53	608×608		56	43.5	26.7	2020
SAPD[31]	X 101-32 x 4 d-DCN	800×800	√	9	46.6	27.3	2020
PAA[38]	R 101-DCN	1 333×800*	√	13	47.4	27.9	2020
GFLV1[39]	R101-DCN	1 333×800	√	15	47.3	28.0	2020
BorderDet[36]	X 101-64 x 4 d	1 333×800*	√	8	47.2	28.1	2020
Scaled-YOLOv4[40]	CD 53 s	640×640		65	47.5	28.2	2021
RepPointsV2[35]	R 101-DCN	1 333×800	√	10	48.1	28.7	2020
YOLOX-L[32]	CSPv 5	640×640		69	49.0	29.6	2021
GFLV2[34]	X 101-32 x 4 d-DCN	1 333×800	√	11	49.0	29.7	2021
PP-YOLOv2[33]	R 50-vd-DCN	608×608		68	49.4	29.8	2021
SODet	CoHiT-S	608×608		40	42.0	27.3	
SODet	CoHiT-M	608×608		37	44.1	27.9	
SODet	CoHiT-L	608×608		33	46.1	28.4	
SODet	CoHiT-X	608×608		26	47.9	29.5	
SODet	CoHiT-X	800×800		17	49.4	31.5	

速度比SAPD快将近4倍. 相同输入尺寸下,主干网络为CoHiT-X时AP和AP_S分别为47.9%和29.5%,检测精度接近于YOLOX-L[32]、PP-YOLOv 2[33]等算法,同时在精度相当的条件下速度为GFLV 2[34]算法的1.5倍. 以CoHiT-L为主干网络的SODet算法与具有更大的输入尺寸且采用多尺度训练策略的RepPoingtsV 2[35]和BorderDet[36]等算法检测性能接近,同时速度为其3倍. 在800×800的输入图像尺寸时,以CoHiT-X为主干网络的SODet算法AP_S达到了31.5%,同时具有17 FPS的检测速度. 实验结果表明,SODet算法在具有较高的小目标检测精度的同时具有较快的检测速度,很好地权衡了精度和速度.

3 结 论

本文针对现有小目标检测算法对小目标检测精度低等问题,提出了一种基于Transformer和CNN的小目标检测算法SODet. 在主干网络CoHiT中利用Transformer和CNN捕获全局和局部信息,并利用自适应特征选择模块进行融合;在特征融合网络FrtPN中融合主干网络输出的额外尺度特征图,针对小目标检测提出大目标抑制单元,在约束大目标特征表达的同时转移小目标特征;在损失函数设计时,利用EIOU损失函数计算包围框回归误差,同时加速收敛,并利用Focal loss将损失函数进行改进,提升小目标的训练效果. SODet算法在MS COCO数据集达到了31.5%的小目标检测精度,且与其他主流方法相比具有较强的竞争力,实验结果表明,所提出方法能够有效地提升小目标检测精度,同时具备一定实时检测能力.

参考文献(References)

[1] Luo X F, Hu H F. Selected and refined local attention module for object detection[J]. Electronics Letters, 2020, 56(14): 712-714.

[2] Masita K L, Hasan A, Shongwe T. Deep learning in object detection: A review[C]. International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems. Durban, 2020: 1-11.

[3] Zheng Q Y, Chen Y. Feature pyramid of bi-directional stepped concatenation for small object detection[J]. Multimedia Tools and Applications, 2021, 80(13): 20283-20305.

[4] Liu G, Han J, Rong W Z. Feedback-driven loss function for small object detection[J]. Image and Vision Computing, 2021, 111: 104197.

[5] Qu J S, Su C, Zhang Z W, et al. Dilated convolution and feature fusion SSD network for small object detection in remote sensing images[J]. IEEE Access, 2020, 8: 82832-82843.

[6] 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.

[7] Redmon J, Farhadi A. Yolov 3: An incremental improvement[J/OL]. 2018, arXiv: 1804.02767.

[8] Liu W, Anguelov D, Erhan D, et al. SSD: Single shot multi box detector[C]. European Conference on Computer Vision, 2016, 9905: 21-37.

[9] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(6): 318-327.

[10] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[J]. IEEE Conference on Computer Vision and Pattern Recognition: CVPR, 2016: 770-778.

[11] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision[J]. IEEE Conference on Computer Vision and Pattern Recognition: CVPR, 2016: 2818-2826.

[12] Yu Fisher, Koltun V. Multi-scale context aggregation by dilated convolutions[J/OL]. 2015, arXiv: 1511.07122.

[13] 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 and Machine Intelligence, 2015, 37(9): 1904-1916.

[14] Hu J, Shen L, Albanie S, et al. Squeeze-and-excitation networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(8): 2011-2023.

[15] Woo, S, Park J, Lee J Y, et al. CBAM: Convolutional block attention module[C]. Proceedings of the European Conference on Compute Rvision. Munich, 2018, 11211: 3-19.

[16] Wang Q L, Wu B G, Zhu P F, et al. ECA-net: Efficient channel attention for deep convolutional neural networks[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, 2020: 11531-11539.

[17] Li X, Wang W H, Hu X L, et al. Selective kernel networks[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, 2019: 510-519.

[18] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J/OL]. 2017, arXiv: 1706.03762.

[19] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: Transformers for image recognition at scale[J/OL]. 2020, arXiv: 2010.11929.

[20] Liu Z, Lin Y T, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]. IEEE/CVF International Conference on Computer Vision. Montreal, 2021: 9992-10002.

[21] Zhang Q L, Yang Y B. ResT: An Efficient transformer for visual recognition[J/OL]. 2021, arXiv: 2105.13677.

[22] Zhang Y F, Ren W Q, Zhang Z, et al. Focal and efficient IOU loss for accurate bounding box regression[J/OL]. 2021, arXiv: 2101.08158.

[23] Bochkovskiy A, Wang C Y, Liao H Y M. YOLOv 4: Optimal speed and accuracy of object detection[J/OL]. 2020, arXiv: 2004.10934.

[24] Liu S, Qi L, Qin H F, et al. Path aggregation network for instance segmentation[J]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018: 8759-8768.

[25] Ghiasi G, Lin T Y, Le Q V. NAS-FPN: Learning scalable feature pyramid architecture for object detection[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, 2019: 7029-7038.

[26] Loshchilov I, Hutter F. Decoupled weight decay regularization[J/OL]. 2017, arXiv: 1711.05101.

[27] He K M, Gkioxari G, Dollár P, et al. Mask R-CNN[C]. IEEE International Conference on Computer Vision. Venice, 2017: 2980-2988.

[28] Peng Z L, Huang W, Gu S Z, et al. Conformer: Local features coupling global representations for visual recognition[C]. IEEE/CVF International Conference on Computer Vision. Montreal, 2021: 357-366.

[29] Mao M Y, Zhang R R, Zheng H H, et al. Dual-stream network for visual recognition[J/OL]. 2021, arXiv: 2105.14734.

[30] Wang W H, Xie E Z, Li X, et al. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions[C]. IEEE/CVF International Conference on Computer Vision. Montreal, 2021: 548-558.

[31] Zhu C C, Chen F, Shen Z Q, et al. Soft anchor-point object detection[J/OL]. 2020, arXiv: 1911.12448.

[32] Ge Z, Liu S T, Wang F, et al. YOLOX: Exceeding YOLO series in 2021[J/OL]. 2021, arXiv: 2107.08430.

[33] Huang X, Wang X X, Lv W Y, et al. PP-YOLOv2: A practical object detector[J/OL]. 2021, arXiv: 2104.10419.

[34] Li X, Wang W H, Hu X L, et al. Generalized focal loss V 2: Learning reliable localization quality estimation for dense object detection[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, 2021: 11627-11636.

[35] Chen Y H, Zhang Z, Cao Y, et al. RepPoints V 2: Verification meets regression for object detection[J/OL]. 2020, arXiv: 2007.08508.

[36] Qiu H, Ma Y C, Li Z M, et al. BorderDet: Border feature for dense object detection[C]. Computer Vision — ECCV, 2020, 12346: 549−564.

[37] Zhang S F, Chi C, Yao Y Q, et al. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection[J]. IEEE/CVF Conference on Computer Vision and Pattern Recognition: CVPR, 2020: 9756-9765.

[38] Kim K, Lee H S. Probabilistic anchor assignment with IoU prediction for object detection[C]. Computer Vision — ECCV, 2020, 12370: 355−371.

[39] Li X, Wang W H, Wu L J, et al. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection[J/OL]. 2020, arXiv: 2006.04388.

[40] Wang C Y, Bochkovskiy A, Liao H Y M. Scaled-YOLOv 4: Scaling cross stage partial network[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, 2021: 13024-13033.

作者简介

赵亮 (1980−), 男, 教授, 博士, 从事智能检测、机器视觉等研究, E-mail: zhaoliang@xauat.edu.cn;

刘世鹏 (1997−), 男, 硕士生, 从事计算机视觉、图像处理的研究, E-mail: hwpq5658@163.com.

(责任编辑：魏 冰)