计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2024,33(2):176−187 [doi: 10.15888/j.cnki.csa.009387]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

融合多层次浅层信息的航拍小目标检测①

秦云飞[1], 崔晓龙[2], 程　林[1], 樊继东[1]

[1](湖北汽车工业学院 汽车工程学院, 十堰 442002)
[2](中南民族大学 计算机科学学院, 武汉 430074)
通信作者: 程　林, E-mail: 1721140082@qq.com

摘　要: 针对小目标检测及目标被遮挡的问题, 本文基于 VisDrone2019 数据集构建相应交通场景, 提出一种小目标检测算法. 首先, 充分利用主干网络的浅层特征改善小目标漏检的问题, 通过在 YOLOv7 算法原有的网络结构上增加小目标检测层 P2, 并在 P2 小目标检测层的模型上为特征融合网络添加多层次浅层信息融合模块, 从而提高算法小目标检测效果. 其次, 使用全局上下文模块构建目标与全局上下文的联系, 增强模型区分目标与背景的能力, 改善目标因遮挡而出现特征缺失情况下的被检测效果. 最后, 本文采用专为小目标设计的损失函数 *NWD* 代替基线模型中的 CIoU 损失函数, 从而解决了 IoU 本身及其扩展对微小物体的位置偏差非常敏感的问题. 实验表明, 改进后的 YOLOv7 模型在航拍小目标数据集 VisDrone2019 (测试集和验证集) 上面 *mAP*.5:.95 分别有 2.3% 和 2.8% 的提升, 取得了十分优异的检测效果.

关键词: 浅层特征; 全局上下文模块; *NWD* 损失函数; 小目标检测; 特征融合; 目标检测

引用格式: 秦云飞,崔晓龙,程林,樊继东.融合多层次浅层信息的航拍小目标检测.计算机系统应用,2024,33(2):176–187. http://www.c-s-a.org.cn/1003-3254/9387.html

Small Target Detection for Aerial Photography Fusing Multi-layer Shallow Information

QIN Yun-Fei[1], CUI Xiao-Long[2], CHENG Lin[1], FAN Ji-Dong[1]

[1](Information School of Automotive Engineering, Hubei University of Automotive Technology, Shiyan 442002, China)
[2](College of Computer Science, South-central Minzu University, Wuhan 430074, China)

Abstract: To solve the problem of small target detection and target occlusion, this study constructs corresponding traffic scenes based on the VisDrone2019 data set and proposes a small target detection algorithm. First, the shallow features of the backbone network are fully used to improve the problem of missing small targets. The small target detection layer P2 is added to the original network structure of the YOLOv7 algorithm, and a multi-level shallow information fusion module is added to the feature fusion network of the model of the small target detection layer P2, so as to improve the small target detection effect of the algorithm. Secondly, the global context module is used to build the connection between the target and the global context, enhance the ability of the model to distinguish between the target and the background, and improve the detection effect when the target is missing features due to occlusion. Finally, the CIoU loss function in the baseline model is replaced by *NWD*, a loss function specially designed for small targets in this study, so as to solve the problem that IoU itself and its extension are highly sensitive to the position deviation of small targets. Experiments show that the improved YOLOv7 model has improved by 2.3% and 2.8% respectively in the small target aerial photography data set VisDrone2019 (test set and validation set) with *mAP*.5:.95, achieving excellent detection results.

Key words: shallow feature; global context module; *NWD* loss function; small target detection; feature fusion; target detection

① 收稿时间: 2023-05-27; 修改时间: 2023-06-26, 2023-08-08; 采用时间: 2023-09-18; csa 在线出版时间: 2024-01-02
CNKI 网络首发时间: 2024-01-03

近年来, 随着无人机装备技术的飞跃发展, 无人机不仅在军用领域发挥着重要作用, 同样在民用领域也得到了广泛的应用, 无人机与摄像头监控摄像头相比, 可以远距离对道路交通情况进行拍摄, 提供更全面的道路环境, 无人机提供的图像中目标车辆较小, 容易出现漏检. 其次, 无人机在多个角度对复杂交通进行拍摄, 导致目标车辆之间出现相互遮挡, 造成检测精度低的后果. 因此改善小目标与被遮挡目标的检测效果, 对提高算法的实用性很有帮助.

针对小目标检测精度低的难题, 国内外众多学者分别从多尺度预测、基于上下文信息、注意力机制、小目标数据增强和提高特征分辨率等角度展开工作.

单层特征对小目标的表征能力太差, 故采用高层特征与低层特征融合的多尺度预测策略. 陈欣等[1]设计了一种特征融合机制, 将分辨率高的浅层特征图与具有丰富语义信息的深层特征图进行融合, 在特征图之间构建特征金字塔, 从而对小目标特征进行增强, 以达到提升小目标检测能力的目的. 李凯等[2]为了充分利用低层特征图的高分辨率和深层特征图的高语义性, 提出了一种多尺度信息融合 (MSIF) 的特征提取方法, 通过统一通道数和上采样操作, 将不同特征图的信息融合起来, 加强层与层之间的联系, 从而提升小目标的检测性能.

针对小目标携带的特征信息有限, 引入目标的上下文信息, 可以挖掘目标与周围目标和环境的联系, 提高对目标特征的判别能力, CoupletNet[3]使用 PSRoI 池化分支来捕获目标的局部信息和 RoI 池化分支来编码全局和上下文信息, 最后使用 1×1 卷积和逐元素相加的方式将两个分支的特征融合起来. 这使得模型能够缓解因缺乏上下文信息而导致目标漏检的问题. 另一种引入上下文信息的方式将部分卷积模块替换成 Transformer 模块, 计算特征图内所有像素之间的联系. 祝星旭[4]使用 Transformer 模块代替 YOLOv5 主干特征提取网络中最底部的模块和特征融合网络中的部分模块, 并增加一个只有 4 倍采样率的低分辨率的分支送入特征融合网络, 增加对微小物体的检测能力, 在小目标数据集上的结果显示, 检测精度得到了很好的提升.

小目标通常缺乏空间维度的区分性特征信息, 引入拥有空间注意力模块可以改善网络对目标位置的敏感度. 李子豪等[5]使用自适应协同注意力机制 ACAM, 避免全通道输入特征在输出单通道的空间注意力权重过程中不同通道特征融合后大量细节信息丢失问题, 使用多局部的方法保留更精细的空间注意力权重增加空间信息区分度从而提升有效特征利用率.

小目标数据增强通常作为数据预处理阶段, 通过增加小目标数量完成. Kisantal 等[6]对包含小物体的图像进行过采样与执行小物体增强, 促使模型更多地关注小物体, 其中将小目标副本粘贴到不同的位置, 既增加了小目标的数量, 又进一步挖掘了目标和周围像素的内在联系, 增加小目标背景的泛化能力, 增强小目标检测的鲁棒性. 郭磊等[7]受 Mosaic 思想的启发, 采用 Mosaic 方法的增强版—Mosaic-8, 即采用 8 张图片随机裁剪、随机排列、随机缩放, 然后组合成一张图片, 以此来增加样本的数据量, 同时合理引入一些随机噪声, 增强网络模型对图像中小目标样本的区分力, 提升了模型的泛化力.

与上述策略不同, 本文以充分利用主干网络浅层特征的思想为指导, 首先使用小目标检测层和多层次浅层信息模块相结合的策略, 解决航拍角度下小目标检测精度低的问题, 增加小目标检测层在图 1 中用红字标出, 在多层次浅层信息融合模块在图 1 中用 adaptive fusion 表示. 使用全局上下文 (GC) 模块改善被遮挡目标的检测效果和 *NWD* 损失函数[8]提升小目标检测精度, GC 模块加在图 1 中的 P2、P3 层经过 ELAN 结构提取特征之后. 具体如下.

(1) 在 YOLOv7 算法[9]的网络结构上增加小目标检测层 P2, 该检测层与基线模型中 3 个尺度的检测层相比, 它的感受野面积与小目标面积的重叠度最大, 且该检测层的特征大多数来源于主干网络的浅层特征. 因此, 该检测层能够充分提取小目标准确的对象描述信息, 从而提升小目标的检测精度.

(2) 在 P2 小目标检测层的基础上为特征融合网络添加多层次浅层信息融合模块[10], 进一步充分利用主干网络的浅层特征信息, 从而提高算法对小目标的检测效果.

(3) 使用全局上下文模块[11]构建目标与全局上下文的联系, 增强模型区分目标与背景的能力, 改善模型对特征缺失目标 (被遮挡导致) 的检测效果, 使得模型更关注重点区域, 而忽略不必要的区域.

(4) 采用专为小目标设计的损失函数 *NWD* 代替基线模型中的 CIoU 损失函数[12], 从而解决了 IoU 本身及其扩展对微小物体位置偏差非常敏感的问题.

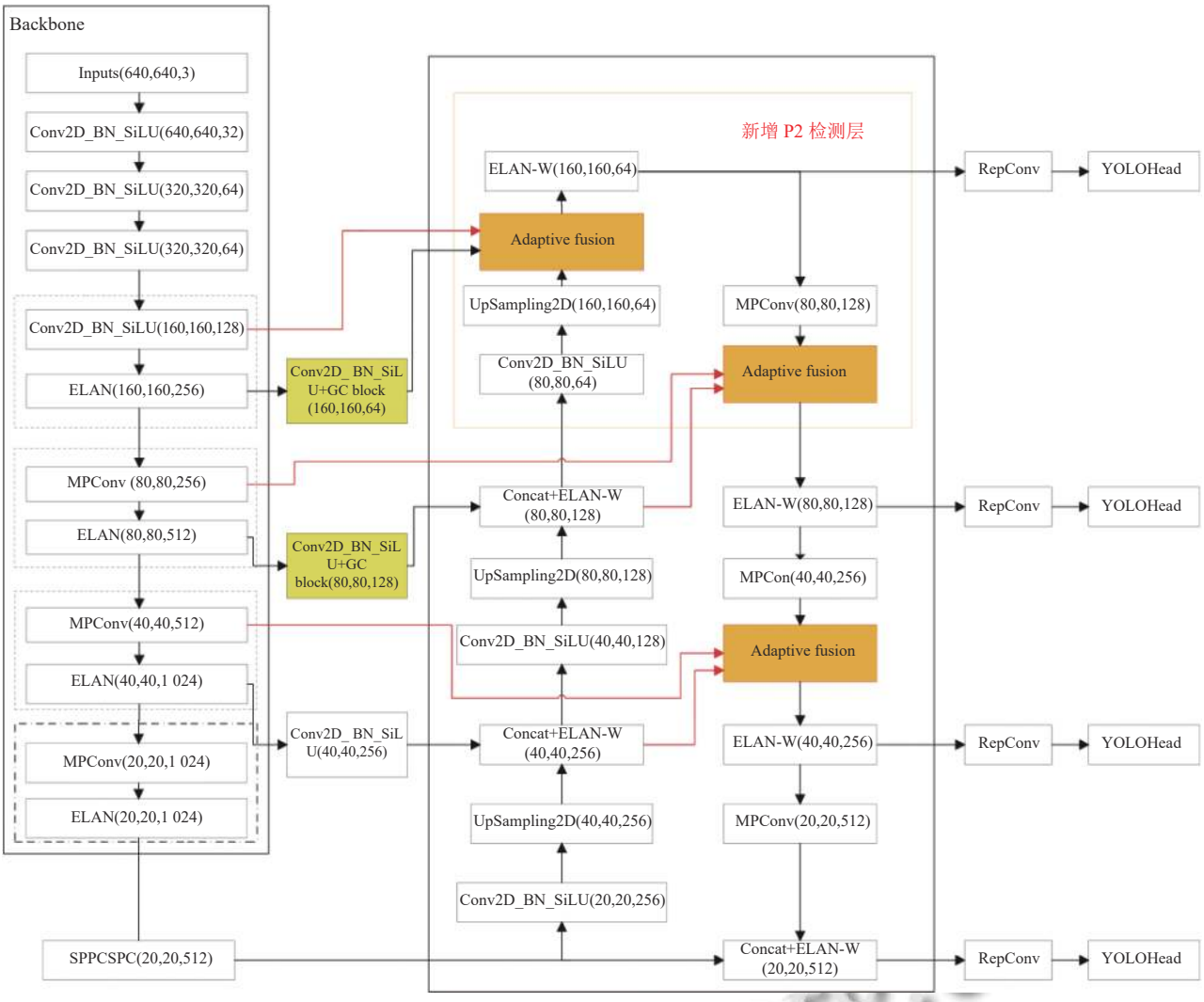


图 1 融合多层次浅层信息的小目标检测模型

1 YOLOv7 算法改进

1.1 小目标检测层

在卷积神经网络中卷积层的不断堆叠会扩大感受野的范围, 且感受野之间重叠区域的增加, 导致图像信息被进一步压缩. 因此, 网络更容易获取图像整体性的一些信息. 这种特征提取策略是有利于网络检测大尺度的目标, 因为大尺度目标在不断下采样的过程中, 特征信息不会随着图像的压缩而消失. 与大尺度不同, 小尺度目标本身所含像素面积较小, 且面积会随着图像压缩尺寸的增加而逐渐降低, 甚至完全被丢失. 例如 16×16 的小目标, 在输入图像尺寸为 640×640 的网络中, 图片被下采样到 40×40 的特征图中只剩 1×1 的尺寸. 而在大多数目标检测模型中, 最低会下采样到 20×20 的尺寸. 因此, 容易出现过采样导致小目标信息丢失, 造成检测精度低, 甚至造成目标漏检的情况发生. 为了满足小目标检测任务, 本文增加一个小目标检测层 P2, 该检测层分辨率为 160×160, 如图 1 所示. P2 特征层的特征大多数来自主干网络的底层特征, 底层特征只经过少量卷积层的堆叠, 保留了大量有利于定位的细粒度信息 (例如边缘、轮廓、纹理等). 浅层特征层中小目标的信息也没有丢失, 因此可以改善小目标的检测性能. 如图 2 所示, 低层特征能够描述目标细粒度信息, 而高层特征感受野映射的范围更大, 更加注重被检测图片中大范围的信息, 需检测的小目标面积所占比例十分小.

1.2 多层次浅层信息融合模块

在基线模型上添加的小目标检测层 P2 中的大多数特征来自模型的浅层特征. 添加小目标检测层后的

模型, 在官方数据集 VisDrone2019 (测试集和验证集) 上 *mAP*.5:.95 分别提升了 1.5% 和 1.1%, 效果十分显著. 本节正是基于该角度出发, 向特征融合网络中添加多层次浅层信息融合模块. 该模块将自主干网格特征层 P2、P3、P4 的上一层浅层特征作为其中一个分支, 由于 P5 层的上一层特征层已经属于高级语义特征层, 包含小目标信息少, 且通道数量较为庞大, 计算成本十分昂贵, 因此只添加 P2、P3、P4 特征层的上一层特征. 上一层特征层分别与对应的 P2、P3、P4 特征层 (P3 和 P4 特征层经过 PAN 中上采样堆叠后, 使用 ELAN 提取特征) 和经过 PAN 上下采样的特征层进行多层次浅层信息融合. 该模块将浅层、中层与深层特征进行融合, 以达到充分利用主干网络浅层特征的目的. 相对于 PAN 中原有的特征融合模块, 该模块保留了大量的低层次特征, 有利于提升小目标的检测性能. 图 3 描述了多层次浅层信息融合模块的简图.

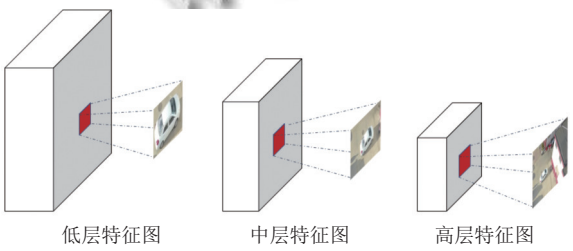


图 2 不同尺度特征图

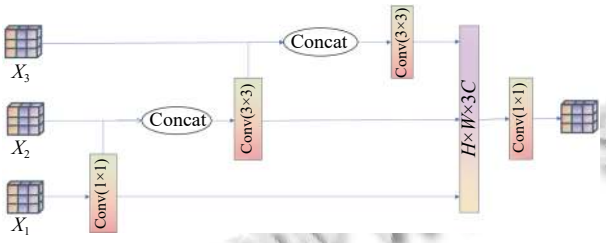


图 3 多层次浅层信息融合模块

模块的输入分别来自主干网络的浅层特征 (X_1)、中层特征 (X_2) 以及来自特征融合网络的深层特征 (X_3). 浅层特征调整通道后参与最后的拼接, 这条通路可以充分利用浅层信息. 中层特征和调整通道后的浅层特征进行融合, 融合后的一个分支向深层特征传递, 一个分支参与最后的拼接. 深层特征融合后经过 3×3 卷积直接参与最后的拼接. 最后将拼接后的特征图利用 1×1 卷积进行通道降维, 如式 (4) 所示:

$$W_1 = \mathrm{Conv}_{1\times1}(X_1) \tag{1}$$

$$W_2 = \mathrm{Conv}_{3\times3}(W_1 + W_2) \tag{2}$$

$$W_3 = \mathrm{Conv}_{3\times3}(W_2 + X_3) \tag{3}$$

$$W = \mathrm{Conv}_{1\times1}(W_1 : W_2 : W_3) \tag{4}$$

其中, $\mathrm{Conv}_{3\times3}$ 和$\mathrm{Conv}_{1\times1}$ 分别表示 1×1 卷积和 3×3 卷积操作, + 号表示Concat堆叠操作. 多层次浅层信息融合模块通过向特征融合网络中增加浅层特征, 以达到提高小目标检测效果的目的.

1.3 *NWD* 损失函数

基于 Anchor 框的卷积神经网络检测算法大多数使用 IoU 及其扩展作为 box 分支的损失函数. 例如 YOLOv7 算法中使用 CIoU 损失函数. 但 IoU 本身及其扩展对微小物体的位置偏差非常敏感, 这使得模型在反向传播过程中更新参数的能力大打折扣. 为了缓解这种情况, 本文使用专为微小目标设计的 *NWD* 损失函数代替 CIoU 损失函数. *NWD* 损失函数由以下几个步骤组成.

(1) 将边界框构建为二维的高斯分布, 并利用两个二维高斯分布之间的距离得出两个 box 框之间的相似性. 二维高斯分布的概率密度函数如式 (5) 所示:

$$f(x|\mu,\Sigma) = \frac{\exp\left(-\frac{1}{2}(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)\right)}{2\pi|\Sigma|^{1/2}} \tag{5}$$

其中, x, μ, Σ分别代表坐标、均值向量和高斯分布的协方差矩阵.

$$(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu) = 1 \tag{6}$$

$$\frac{(x-\mu_x)^2}{\sigma_x{}^2} + \frac{(y-\mu_y)^2}{\sigma_y{}^2} = 1 \tag{7}$$

式 (7) 为椭圆方程, 其中(μ_x,μ_y)是椭圆的中心坐标, σ_x、σ_y是 x 和 y 轴的半轴长度. 当式 (6) 成立时, 式 (7) 中的椭圆将是二维高斯分布的密度轮廓. 因此, 水平边界框$R=(C_x,C_y,w,h)$可以建模为二维高斯分布 $N(\mu,\Sigma)$, 其中(C_x,C_y)表示水平边界框中心点坐标, w和 h分别表示边界框宽和高. 二维高斯分布$N(\mu,\Sigma)$有:

$$\mu = \begin{bmatrix} C_x \\ C_y \end{bmatrix},\ \Sigma = \begin{bmatrix} \frac{w^2}{4} & 0 \\ 0 & \frac{h^2}{4} \end{bmatrix} \tag{8}$$

此外, 边界框 A 和 B 之间的相似性可以转化为两个高斯分布之间的分布距离.

(2) 在该步骤使用最优运输理论中的 Wasserstein distance[13]来计算分布距离. 对于两个二维高斯分布 $\mu_1=N(m_1,\Sigma_1)$ 和 $\mu_2=N(m_2,\Sigma_2)$, μ_1 和 μ_2 之间的二阶 Wasserstein distance 定义为:

$$W_2^2(\mu_1,\mu_2)=\|m_1-m_2\|_2^2+\left\|\Sigma_1^{1/2}-\Sigma_2^{1/2}\right\|_F^2 \tag{9}$$

其中, $\|\cdot\|_F$ 是 Frobenius 规范. 此外, 对于两个包围框 box1=(cx_a,cy_a,w_a,h_a) 和 box2=(cx_b,cy_b,w_b,h_b) 所建模的高斯分布 N_a 和 N_b, 式 (9) 能被进一步简化:

$$W_2^2(N_a,N_b)=\left\|\left(\left[cx_a,cy_a,\frac{w_a}{2},\frac{h_a}{2}\right]^{\mathrm{T}},\left[cx_b,cy_b,\frac{w_b}{2},\frac{h_b}{2}\right]^{\mathrm{T}}\right)\right\|_2^2 \tag{10}$$

然而, $W_2^2(N_a,N_b)$ 是距离度量, 不能直接作为相似性度量 (IoU 值在 0–1 之间). 因此使用指数归一化的形式, 获得最后的 *NWD* 度量, 如式 (11) 所示:

$$NWD(N_a,N_b)=\exp\left(-\frac{\sqrt{W_2^2(N_a,N_b)}}{C}\right) \tag{11}$$

其中, C 是一定范围内稳健的常数. *NWD* 损失函数与 IoU 损失函数相比, 在检测微小目标最主要的优点有: (1) 位置偏差的光滑性, 相对于 IoU 系列损失函数对微小位置偏差的敏感性, 位置偏差导致的 *NWD* 变化更为平滑, 更好的平滑性是有利于预测框定位. (2) 能够测量不重叠或相互包含的包围盒之间的相似性. 使用该损失函数代替 CIoU 损失函数更符合小目标检测的任务需求.

1.4 GC 全局上下文模块

YOLOv7 模型由大量的卷积层以一定的规则堆叠而成, 而卷积运算是将局部区域的特征进行加权输出, 感受野仅限于窗口大小, 例如 3×3, 5×5 大小等. 为了捕获长距离远程依赖关系, 卷积神经网络依靠反复的堆叠形成更大范围的感受野来建立[14].

如图 4 所示, 在卷积运算中是否能检测到篮球特征仅由局部区域决定, 而与周围场景毫无相关. 但篮球和人之间存在重要的依赖关系, 由于人的特征距离较远, 且浅层特征层感受野较小. 因此, 在浅层的卷积运算中被忽视掉, 只有通过不断堆叠增大篮球所在区域的感受野才能获得远距离人的特征. 因此, 卷积运算中缺乏主动构建远程依赖关系的模块. 且浅层特征的感受野较小, 因此构建远程依赖尤为重要. 本文增加的小目标检测层里浅层特征占据了大多数, 因此, 利用好远程依赖关系能够提升小目标的检测性能.

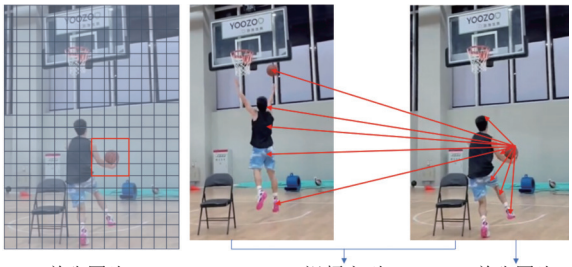


图 4 卷积运算和非局部运算

如图 5 所示, 全局上下文模块 (GC block) 由被简化的非局部均值模块 (simplified NL block)、压缩激励模块 (SE block) 组合而成, 其组成过程如下 3 个步骤所示.

(1) 非局部均值模块

Wang 等[15]提出一种建立远距离依赖的可嵌入式运算块: 非局部均值模块. 该运算块将某个位置的响应计算为所有位置特征的加权和. 非局部运算的定义如式 (12) 所示:

$$y_i=\frac{1}{C(x)}\sum_{\forall j}f\left(x_i,x_j\right)g\left(x_j\right) \tag{12}$$

其中, i 是响应的输出位置的索引, j 是枚举所有可能位置的索引, x 是输入特征, y 是与 x 相同大小的输出特征. 一元函数 g 将 j 位置的输入信号进行映射, 位置响应通过 $C(x)$ 进行归一化. 为了简单起见, g 为线性函数, g 函数如式 (13) 所示:

$$g\left(x_j\right)=W_g x_j \tag{13}$$

其中, W_g 是可学习权重矩阵. f 函数是计算 i 与所有 j 之间的标量 (表示两者的密切关系), 最常用的 f 函数为高斯函数的扩展, 如式 (14) 所示:

$$f\left(x_i,x_j\right)=\mathrm{e}^{\theta(x_i)^{\mathrm{T}}\phi(x_j)} \tag{14}$$

该函数在嵌入式空间中计算二者的相似性, 其中 $\theta(x_i)=W_\theta x_i$ 和 $\phi\left(x_j\right)=W_\phi x_j$, 同时 $C(x)$ 将被定义为 $C(x)=\sum_{\forall j}f(x_i,x_j)$, W_θ 和 W_ϕ 为可学习的权重矩阵.

将非局部运算块包装成一个可嵌入其他架构的非局部运算块, 如式 (15) 所示:

$$z_i=W_z y_i+x_i \tag{15}$$

其中, y_i 如式 (12) 所示, $+x_i$ 表示残差连接, W_z 为权重矩

阵. 残差连接允许非局部块插入任何预训练模型中, 而不破坏其初始行为 (例如将 W_z 矩阵初始化为 0 即可).

与卷积运算的局部性渐进行为不同, 非局部运算块无视两个位置之间距离, 可以计算任意两个位置之间的交互, 从而捕获远程依赖关系. 非局部运算块也很容易与卷积层相结合, 从而构建非局部和局部信息相结合的多信息结构块. 但由于其为特征图中的每个位置计算注意力图, 非局部块的时间和空间复杂度为查询位置和的二次方. 其庞大的计算量, 导致该模块在其他模型架构的底层特征中不能很好的应用.

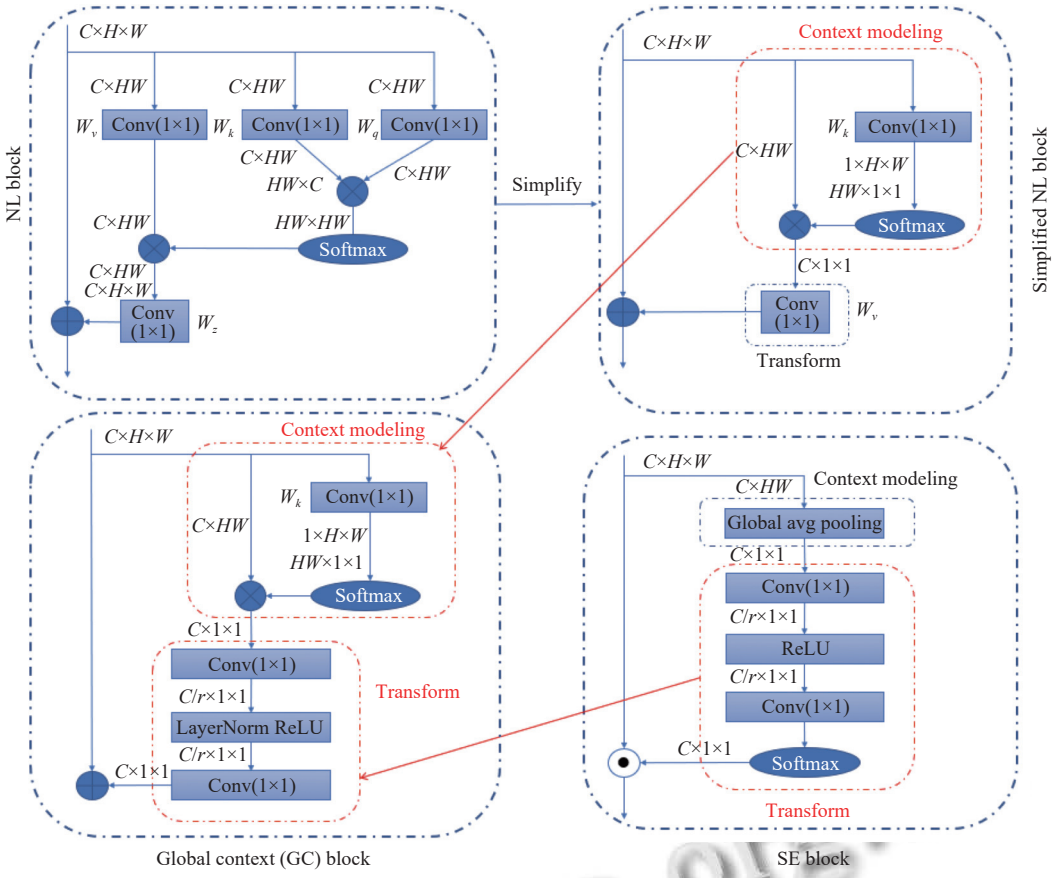


图 5 训练集目标被遮挡情况

(2) 简化的非局部均值模块

2019 年, Cao 等[11]通过可视化不同查询位置的非局部均值模块产生的热力图, 发现不同查询位置的热力图几乎相同. 其次统计不同查询位置的全局上下文之间的距离, 发现不同上下文之间的距离远比输入特征中不同位置之间的距离小. 表明非局部块建模的上下文对于不同查询位置的效果一样, 因此将查询分支移除. 将非局部运算块进行了简化, 计算的全局注意力图为所有查询位置共享, 节省了大量的计算成本. 被简化的非局部运算块如式 (16) 所示:

$$z_i = x_i + \sum_{j=1}^{N_p} \frac{\exp\left(W_k x_j\right)}{\sum_{m=1}^{N_p} \exp\left(W_k x_m\right)} \left(W_v \cdot x_j\right) \tag{16}$$

由于简化的非局部块中 W_z 的存在与否不影响精度, 因此在公式中移除式 (15) 中的 W_z. 其中 W_k 和 W_v 表示线性变换矩阵, i、j 和 m 为输入特征的位置索引.

为了进一步降低被简化局部块的计算成本, 将 W_v 移出. 变换后如式 (17) 所示:

$$z_i = x_i + W_v \sum_{j=1}^{N_p} \frac{\exp\left(W_k x_j\right)}{\sum_{m=1}^{N_p} \exp\left(W_k x_m\right)} x_j \tag{17}$$

W_v 矩阵的计算复杂度由 $\mathrm{O}\left(HWC^2\right)$ 变成 $\mathrm{O}\left(C^2\right)$. 与式 (14) 对比, 式 (17) 的第 2 项中没有 i 下标, 表明该项与查询位置无关, 所有查询位置都共享. 因此直接将全局上下文建模为所有位置的加权平均值, 以获得全局上下文特征.

(3) 引入 SE block

在简化的非局部模块中, 变换模块 (图 5 中 Transform 模块) 的参数量最大, 包含C^2参数量的 1×1 卷积, 若将此局部模块放入高层特征 (P5 层), 高层中的多通道数将使这个模块的参数量剧增. 因此原文中引入轻量型注意力模块 (SE block) 到变换模块, 得到最后的全局上下文模块. 将 1×1 卷积用瓶颈变换模块代替, 可以将参数量由C^2变成$2\times C\times C/r$, 其中r是瓶颈比. SE block[16] 的引入大大降低了 GC block 计算量. GC block 如式 (18) 所示:

$$z_i = x_i + W_{v2}ReLU\left(LN\left(W_{v1}\sum_{j=1}^{N_p}\frac{\exp\left(W_k x_j\right)}{\sum_{m=1}^{N_p}\exp\left(W_k x_m\right)}x_j\right)\right) \tag{18}$$

其中, $ReLU$ 为非线性激活函数, LN 为层标准化, W_{v2} 为线性变换矩阵. 在瓶颈变换 (在 $ReLU$ 之前) 中添加了层标准化, 能够简化两层瓶颈变换带来的优化困难问题, 以及能够作为一个有利于泛化的正则化器.

如图 5 所示, GC block 中包含 SNL block 中的 context modeling 模块和 SE block 中的 Transform 模块, 最后使用广播逐元素相加操作对输入特征和全局特征进行融合, 得到模块的输出结果.

2 实验及结果分析

2.1 实验环境与参数设置

模型在数据集上面训练 300 轮 (每轮都使用全部的训练集图片数据), 改进前后的模型都使用相同的初始权重文件. 模型中需要固定的超参数: batch_size 在训练和测试阶段均设置为 16, 初始学习率为 0.01, 循环学习率 0.1, 学习率采用 warm-up 方式, 预热的轮数为 3 轮, 预热阶段结束后回到初始学习率 0.01, 预热的动量为 0.8, 随后学习率开始逐渐下降, 防止模型因学习率过大产生过拟合现象. SGD 在较大的数据集上泛化性更强且只需要存储当前的参数和梯度, 存储需求相对 Adam 较低, 因此本文实验采用的优化器为包含动量的随机梯度下降法 (SGD), 优化器中动量值为 0.937, 权重衰减因子为 0.0005.

2.2 数据集解析

本文实验中使用无人机拍摄的小目标数据集 VisDrone2019. 该数据集中大量的小目标和被遮挡目标是算法检测的难点. 训练集有 6471 张图片, 验证集 548 张图片, 测试集 1610 张图片, 总共图片为 8599 张图片. 该数据集一共拥有 12 个类别, 每个类别数量差异很大, 这与现实交通状况类似. 其中 other、ignored regions 类别不考虑[17].

与 COCO2017 将面积小于 32×32 像素的目标定义为小目标不同, 本文更为严格地将小目标分为两个分支, 极小目标 (area≤16×16) 和小目标 (16×16<area≤32×32). 表 1 展示了数据集中每个类别的不同目标大小所占比例, 其中极小目标占比为 26.1%, 小目标占比 34.4%, 二者比例高达 60.5%. 相比于 COCO2017 数据集中小目标所占比例的 41.43%, 可见 VisDrone2019 数据集小目标数量更多. 表 2 为实验环境.

表 1 VisDrone2019 数据集 (训练集) 中目标尺寸分布情况

数据类别	Extre-small (area≤16×16)		Small (16×16<area≤32×32)		Normal (area>32×32)	
	目标个数	目标占比 (%)	目标个数	目标占比 (%)	目标个数	目标占比 (%)
Pedestrian	34379	43.3	30859	38.9	14099	17.8
People	13015	48.1	10495	38.8	3549	13.1
Bicycle	2397	22.9	4706	44.9	3377	32.2
Car	25639	17.7	44217	30.5	75011	51.8
Van	2954	11.8	7801	31.3	14201	56.9
Trunk	1064	8.2	2920	22.7	8891	69.1
Tricycle	659	13.7	1520	31.6	2633	54.7
Awning-tricycle	393	12.1	972	30.0	1881	57.9
Bus	422	7.1	1235	20.9	4269	72.0
Motor	8647	29.2	13325	44.9	7675	25.9
合计	89569	26.1	118050	34.4	135586	39.5

2.3 评价指标

目标检测常用召回率 (*Recall*, *R*)、准确率 (*Precision*, *P*)、均值平均精度 (*mAP*)、模型参数量 (Parameters)、延迟这 5 项评价指标对模型进行评估.

准确率是模型预测中被正确划分为正样本数占全部被划分为正样本数的比例, 如式 (19) 所示:

$$P=\frac{T_P}{T_P+F_P} \tag{19}$$

其中, T_P (true positive) 为被正确识别的正样本数, F_p (false positive) 为被错误识别的正样本数.

表 2 实验环境

名称	环境参数
操作系统	Ubuntu 20.04
CPU	AMD EPYC 7551P
GPU	A5000-24 GB
深度学习框架	PyTorch 1.7.0, CUDA 10.1

召回率是模型在预测中对数据集中正样本被正确识别的比例, 如式 (20) 所示:

$$R=\frac{T_P}{T_P+F_N} \tag{20}$$

其中, F_N (false negative) 表示实际的正样本错误的被检测为负样本的数量.

为了综合评价模型的性能, 需要将二者联合起来分析, 使用平均精度 (average precision, *AP*) 来平衡二者的关系, 它表示某个类别目标的检测精度. 如式 (21) 所示:

$$AP=\int P\mathrm{d}R \tag{21}$$

以上述 *AP* 值为基础就能得到均值平均精度 (mean average precision, *mAP*), 表示多类别 *AP* 值的平均, 该值越高表明模型性能越好. 如式 (22) 所示:

$$mAP=\frac{1}{C}\sum_{j=1}^{C}AP_j \tag{22}$$

其中, AP_j为 j 类目标的 *AP* 值, 对所有目标类别 *AP* 值求和进行平均计算就可以得出 *mAP* 的大小, 其中 *mAP*.5:.95 表示 IoU 从 0.5 开始间隔 0.05 一直取值到 0.95, 然后求出所有类别均值的 *mAP* 值.

2.4 消融实验

采用消融实验验证改进前后算法的性能, 将模型采用相同的预训练权重训练 300 epoch 之后, 使用最优模型权重在测试集和验证集上面进行测试, 共 4 组模型, 从模型的速度和精度方面进行对比.

A 组为基线模型, 输入图片尺寸为 640×640 像素; B 组在 A 组的基础上增加小目标检测头; C 组在 B 组的基础上增加多层次浅层信息融合模块, 进一步充分利用主干网络的浅层特征; D 组在 C 组的基础上使用 *NWD* 损失函数代替 CIoU 损失函数, 弥补 CIoU 损失函数在小目标检测方面的缺陷. 图 6 展示了 4 组模型在训练过程中验证集的 *mAP*.5:.95 的变化情况.

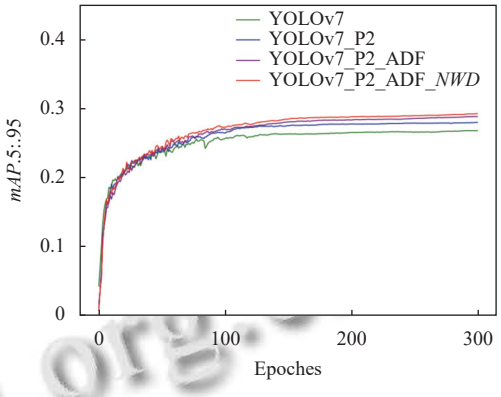


图 6 模型改进前后 *mAP*.5:.95 的比较

由图 6 可知, 向特征融合网络中增加主干网络的浅层特征 (增加小目标检测层和多层次浅层特征融合模块) 对基线模型的检测效果有大幅度的提升, 而 *NWD* 损失函数对模型效果的改进提升较小, 主要是该损失函数仅仅使用在预测框回归的分支.

模型定位损失的实验结果如图 7 所示, 其中定位损失有明显降低的为小目标检测层和 *NWD* 损失函数, 小目标检测层拥有与目标较大的重叠度来精细目标的位置信息, 从而降低模型的位置损失. 而 *NWD* 损失函数对模型定位损失的明显下降, 是由于 *NWD* 损失函数对位置偏差导致的损失函数值的变化更为平滑, 而更好的平滑性是有利于预测框定位损失的.

图 8 展示了 4 组模型每个类别的 *PR* 曲线, 该曲线是在模型训练完成之后, 使用训练过程中训练效果最好的权值文件在验证集上所测试出来的 *mAP*@50 (*mAP*.5). 从图 8 中可知数据集中 4 个数量最多的类别 (car、pedestrian、motor、people) 的 *mAP*@50 均有提升, 而这些类别在实际交通中同样数量最多, 表明 3 种改进策略都十分有效.

表 3 对本文的所有改进策略的在测试集上面的实验结果进行展示, 其中实验 2 在基线模型上增加 P2 小目标检测层, 该检测层大多数的特征来自于主干网络的浅层特征, 改进后模型在测试集上的 *mAP*.5:.95 增加了 1.5%, 而参数量仅增加了 1.6%, 表明通过向特征融合网络中引入包含大量小目标信息的浅层特征的方法, 能够有效地向模型中增加小目标的细粒度信息, 从而大幅度提升模型的检测性能. 实验 3 在实验 2 的基础上增加多层次信息融合模块, 该模块进一步向特征融

合网络中增加主干网络的浅层特征, 以达到充分利用浅层特征改善小目标检测效果的目的. 结果显示, 相较于基线模型, 在测试集上面 *mAP*.5:.95 增加了 1.8%, 而验证集的效果更好, *mAP*.5:.95 增加了 2.0%. 虽然参数量增加较多, 但模型运行速度的增幅却比增加小目标检测层更小, 这得益于浅层特征融合模块, 将原模型中的已存在的模块进行多层次融合, 使得模型运行速度上得到了保证. 实验 4 表示, 在实验 3 的基础上将 CIoU 损失函数改为 *NWD* 损失函数, *mAP*.5:.95 都有小幅度的提升, 该改进策略虽然增加了模型训练成本, 但对模型推理的时间是毫无影响. 因此, 该策略能够无损提升小目标检测精度. 且由图 9 可知, 相较于实验 3, 该策略能够使模型定位损失下降明显, 表明 *NWD* 损失函数能够提高模型的定位能力. 表 3 结果显示, 最终模型的 *mAP*.5:.95 在测试集和验证集上分别提升了 2.3% 和 2.8%.

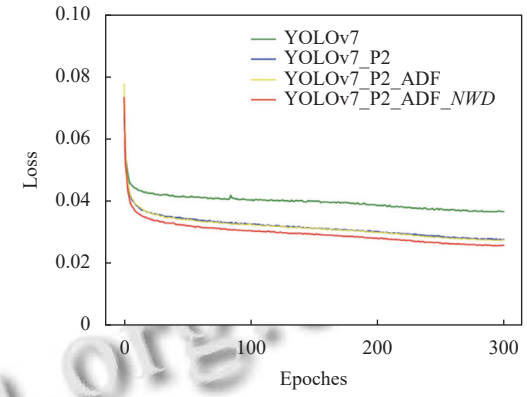
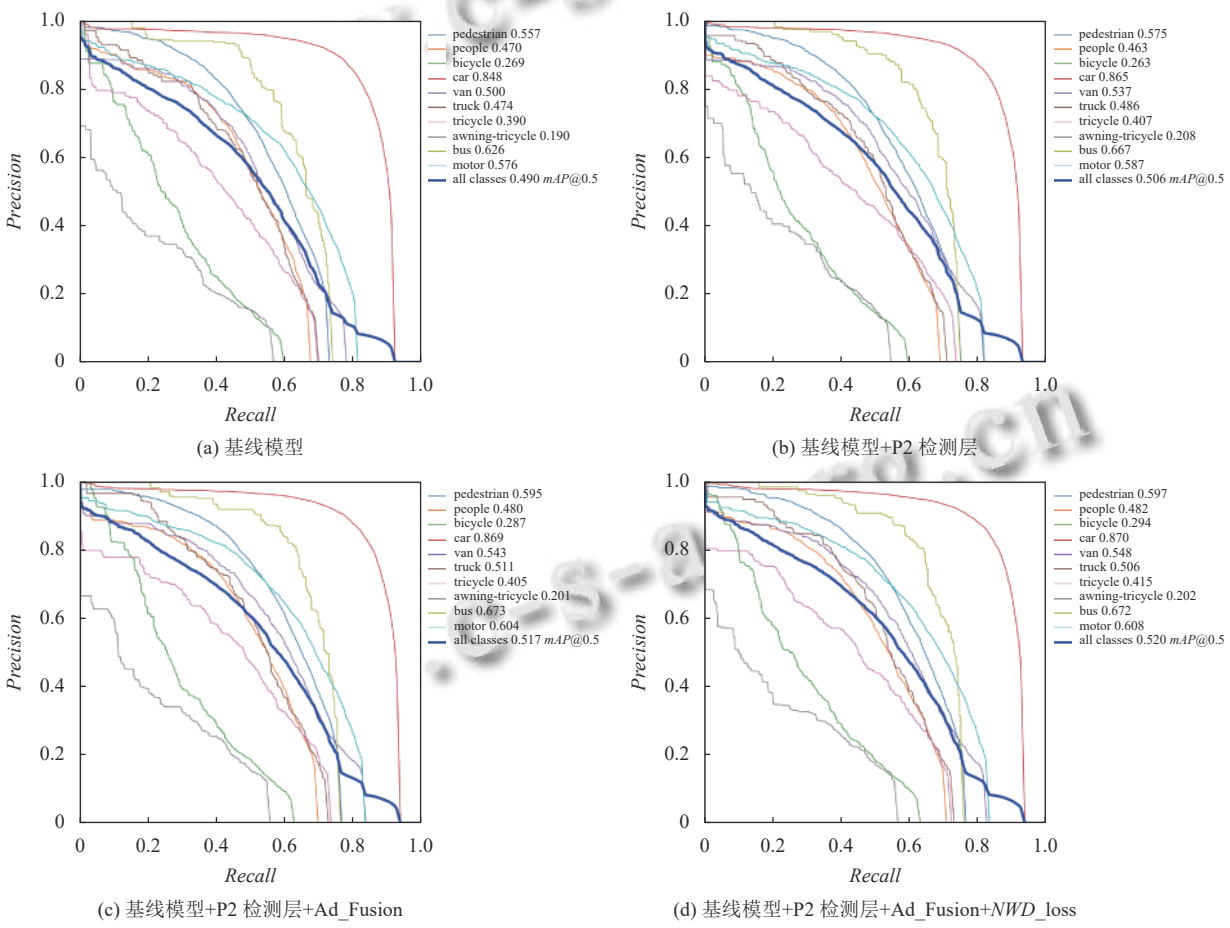


图 7 模型改进前后 Loss 的比较



(a) 基线模型 (b) 基线模型+P2 检测层

(c) 基线模型+P2 检测层+Ad_Fusion (d) 基线模型+P2 检测层+Ad_Fusion+*NWD*_loss

图 8 模型改进前后的 *PR* 曲线

2.5 GC 算法可视化

本文通过 Grad-CAM 算法[18]将加入 GC 模块前后的网络所关注的重点区域进行可视化.

Grad-CAM 算法的工作流程如图 10 所示, 该算法的优点是无需更改网络架构并重新训练, 它就能够将训练好的模型的某一层进行可视化.

GC 全局上下文模块与输入特征的通道数息息相关, 主干网络的深层特征图包含小目标特征信息较少

且通道数较大, 因此只对特征提取网络的浅层 (图 1 所示) 添加全局上下文模块, 以避免过多的加入出现计算量大幅增加而检测精度提升并不明显的情况, 造成计算资源冗余. 如表 3 所示, 增加 GC 全局上下文模块, GC 模块中的 SE block 超参数设置参考文献[16]中将 SE 块中的缩放因子设置为 16, 设置后 VisDrone2019 测试集和验证集上 *mAP*.5:.95 分别有 0.3% 和 0.5% 的提升, 为了展示 GC 模块改进前后的效果进行对比, 将改进前后的模型的 SPPCSP 模块 (该模块位于主干特征提取网络输入到特征融合网络的第 1 层特征层, 模型性能的好坏能够通过它直接的反映) 的特征层进行可视化, 可视化效果如图 11 所示.

表 3 在 VisDrone2019 (测试集和验证集) 上不同模型的性能比较

数据集	实验	基线模型	P2小目标层	Ad_Fusion	*NWD*_loss	GC block	*mAP*.5:.95 (%)	参数量 ($\times 10^6$)	延迟 (ms)
测试集	1	√	—	—	—	—	21.9	36.5	10.6
	2	√	√	—	—	—	23.4 (+1.5)	37.1	12.8
	3	√	√	√	—	—	23.7 (+1.8)	40.8	14.1
	4	√	√	√	√	—	23.9 (+2.0)	40.8	14.1
	5	√	√	√	√	√	24.2 (+2.3)	41.0	16.7
验证集	1	√	—	—	—	—	26.7	36.5	10.6
	2	√	√	—	—	—	27.8 (+1.1)	37.1	12.8
	3	√	√	√	—	—	28.7 (+2.0)	40.8	14.1
	4	√	√	√	√	—	29.0 (+2.3)	40.8	14.1
	5	√	√	√	√	√	29.5 (+2.8)	41.0	16.7

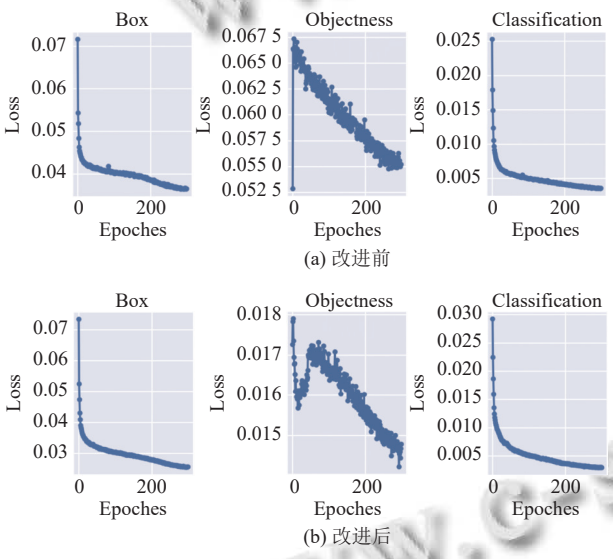


图 9 模型改进前后损失函数变化曲线

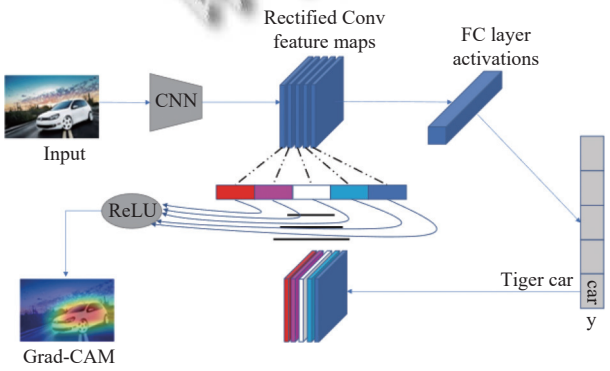
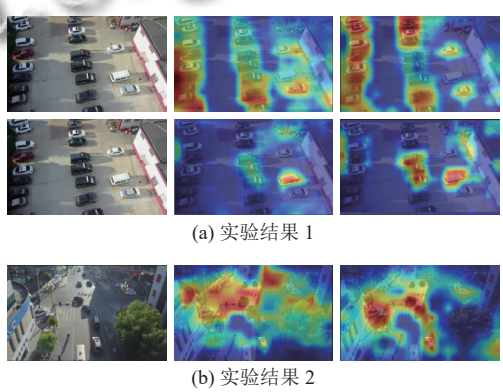


图 10 Grad-CAM 算法工程流程

图 11(a) 中存在小目标, 而图 11(b) 中存在被遮挡目标. 在实验结果 1 中分别展示了 car 和 van 类别下的热力图, 结果 2 为 car 类别的热力图. 在结果 1 (car 类别) 和结果 2 中可以看出, 未添加 GC 模块的模型关注的范围很宽泛, 增加 GC 模块之后, 关注的范围更加精准, 没有目标区域的热度明显变小. 在目标类别数量较少的实验结果 1 (van 类别) 和实验结果 2 中 (car 类别), 增加 GC 模块能够使模型对需关注的目标更聚焦, 从而改善模型区分目标和背景的能力, 从而提高模型对被遮挡目标的检测能力.



(a) 实验结果 1

(b) 实验结果 2

图 11 加入 GC 全局上下文机制后可视化效果

2.6 不同算法对比和检测效果可视化

为了充分展示模型效果, 将改进后的模型与近些年来优秀的一阶段目标检测算法进行对比. 其中 s、m 和 x 为模型的不同子版本. 为了比对效果的真实性,

将实验中的输入图像分辨率统一为 640×640 的大小, 且使用 VisDrone2019 测试集进行实验验证.

如表 4 所示, 在基于航拍角度的特定任务下, 本文提出的模型不仅对基线模型拥有较大的提升, 相对于经典的一阶段 YOLOv5_x 算法模型, *mAP*.5:.95 提升了 4.0%, 相对于原版 YOLOv7 算法模型 *mAP*.5:.95 提升了 2.3%, 并且改进后的模型在延迟上十分有优势. 证明了本文算法在航拍小目标任务的有效性. 图 12 展示了 VisDrone2019 数据集一些检测的可视化结果, 其中包含白天、黑夜、高强度光照和拥挤交通场景下的密集小目标, 结果显示本文所提出的算法能够满足该场景下的小目标检测任务.

表 4 模型性能横向对比

模型	*mAP*.5 (%)	*mAP*.5:.95 (%)	参数量 (M)	延迟 (ms)
YOLOv4	36.3	21.4	63.7	23.8
YOLOv5_s	28.8	15.5	7.6	13.6
YOLOv5_m	32.1	18.2	21.8	18.2
YOLOv5_x	34.6	20.2	87.9	30.3
Our	43.4	24.2	36.9	16.7

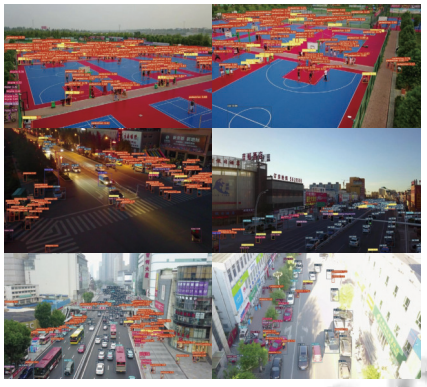


图 12 本文模型航拍角度的检测效果

2.7 模型部署及测试

本文采用的模型部署流程是将 PyTorch 已训练好的 pt 模型文件, 通过 ONNX 扩展库进行转换成通用网络格式 ONNX, 随后使用 TensorRT 进行优化导出 trt 文件. 将输入数据使用 trt 文件进行推理, 得到的结果使用非极大值抑制后进行输出, 得出最终的检查结果. 后续使用时直接加载推理器即可, 这样就可以省去漫长的等待模型优化的过程.

本节对改进前后的算法模型进行部署并测量模型检测速度. 其中 FP_16 表示将模型参数量化为 Float16 进行推理. 其中为了节省推断时间, 在未使用 TensorRT 进行推理加速的模型上使用 model.half() 函数开启半精度. 实验结果如表 5 所示, 其中由于轻量型 GC 全局上下文模块在 TensorRT 中不容易转换, 因此本节不对其使用 TensorRT 进行加速推理.

表 5 不同加速策略下模型的 FPS

模型	FPS
YOLOv7	70
YOLOv7+P2	56
YOLOv7+P2+Ad_Fusion	49
YOLOv7_FP16	224
YOLOv7+P2_FP16	180
YOLOv7+P2+Ad_Fusion_FP16	156

由于每次输入数据后, 得出的 FPS 有少量偏差. 因此, 将 10 次得出的 FPS 取均值. 模型在 TensorRT 优化之后, 推理速度提升了 3 倍以上, 且 3 种不同的模型提升的倍数类似, 这得益于两种改进结构中并没有引入基线模型所没有的模块, 而是充分利用主干的浅层特征所取得的效果. 该实验结果表明将模型量化后使用 TensorRT 加速器在速度上的优良表现.

3 结论

为提高卷积神经网络模型在航拍角度下对小目标的检测能力, 提出一种以增加浅层特征为目的航拍小目标检测算法. 首先, 在 YOLOv7 算法的基础上新增小目标检测层 P2, 结合特征融合网络中的多层次浅层信息融合模块, 充分利用主干网络的浅层特征. 其次, 使用轻量型 GC 全局上下文模块, 以获得更大的感受野和上下文信息, 增强模型区分背景与目标的能力, 提升模型检测被遮挡目标的效果. 最后, 本文使用专为小目标设计的 *NWD* 损失函数代替 CIoU 损失函数, 以缓解 IoU 及其扩展系列的损失函数对小目标微小位置偏差十分敏感的问题. 实验表明, 改进后的模型在官方航拍小目标数据集 VisDrone2019 的测试集和验证集上面 *mAP*.5:.95 分别有 2.3% 和 2.8% 的提升, 取得了十分优异的检测效果.

参考文献

1 陈欣, 万敏杰, 马超, 等. 采用多尺度特征融合 SSD 的遥感图像小目标检测. 光学 精密工程, 2021, 29(11): 2672–2682.
2 李凯, 林宇舜, 吴晓琳, 等. 基于多尺度融合与注意力机制的小目标车辆检测. 浙江大学学报 (工学版), 2022, 56(11): 2241–2250. [doi: 10.3785/j.issn.1008-973X.2022.11.015]

3 Zhu YS, Zhao CY, Wang JQ, *et al*. CoupleNet: Coupling global structure with local parts for object detection. Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017. 4146–4154.
4 祝星顗, 蒋球伟. 基于 CNN 与 Transformer 的无人机图像目标检测研究. 武汉理工大学学报 (信息与管理工程版), 2022, 44(2): 323–331.
5 李子豪, 王正平, 贺云涛. 基于自适应协同注意力机制的航拍密集小目标检测算法. 航空学报, 2023, 44(13): 327944.
6 Kisantal M, Wojna Z, Murawski J, *et al*. Augmentation for small object detection. arXiv:1902.07296, 2019.
7 郭磊, 王邱龙, 薛伟, 等. 基于改进 YOLOv5 的小目标检测算法. 电子科技大学学报, 2022, 51(2): 251–258. [doi: 10.12178/1001-0548.2021235]
8 Wang JW, Xu C, Yang W, *et al*. A normalized Gaussian Wasserstein distance for tiny object detection. arXiv: 2110.13389, 2021.
9 Wang CY, Bochkovskiy A, Liao HYM. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver: IEEE, 2023. 7464–7475.
10 谢椿辉, 吴金明, 徐怀宇. 改进 YOLOv5 的无人机影像小目标检测算法. 计算机工程与应用, 2023, 59(9): 198–206. [doi: 10.3778/j.issn.1002-8331.2212-0336]
11 Cao Y, Xu JR, Lin S, *et al*. GCNet: Non-local networks meet squeeze-excitation networks and beyond. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop. Seoul: IEEE, 2019. 1971–1980.
12 Zheng ZH, Wang P, Ren DW, *et al*. Enhancing geometric factors in model learning and inference for object detection and instance segmentation. IEEE Transactions on Cybernetics, 2022, 52(8): 8574–8586. [doi: 10.1109/TCYB.2021.3095305]
13 Ge Z, Liu ST, Li ZM, *et al*. OTA: Optimal transport assignment for object detection. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 303–312.
14 LeCun Y, Boser B, Denker J, *et al*. Backpropagation applied to handwritten zip code recognition. Neural Computation, 1989, 1(4): 541–551. [doi: 10.1162/neco.1989.1.4.541]
15 Wang XL, Girshick R, Gupta A, *et al*. Non-local neural networks. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7794–7803.
16 Hu J, Shen L, Sun G. Squeeze-and-excitation networks. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7132–7141.
17 黎学飞, 童晶, 陈正鸣, 等. 基于改进 YOLOv5 的小目标检测. 计算机系统应用, 2022, 31(12): 242–250. [doi: 10.15888/j.cnki.csa.008835]
18 Selvaraju RR, Cogswell M, Das A, *et al*. Grad-CAM: Visual explanations from deep networks via gradient-based localization. Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017. 618–626.

(校对责编: 孙君艳)