文章编号:1671-4598(2022)09-0213-08 DOI:10.16526/j.cnki.11-4762/tp.2022.09.032 中图分类号:TP391.41 文献标识码:A

Res2Net 融合注意力机制的 YOLOv4 目标检测算法

张 翔,刘振凯,叶 娜,赵妍祯
(西安建筑科技大学 信息与控制工程学院,西安 710311)

摘要:针对传统目标检测算法容易出现漏检、误检或者有遮挡物时检测困难等问题,提出一种 Res2Net 融合注意力机制的 YOLOv4(Res2Net fusion with attention learning YOLOv4,RFAL YOLOv4)目标检测模型;首先为了获取更多特征图语义信息,通过在一个残差块内构造层次化的类残差连接,引入 Res2Net 替换原 YOLOv4 主干网络中的 ResNet 残差网络结构,可以获取到更细小的特征,同时也增加了模型感受野;其次将 Res2Net 与注意力机制相融合,获取关键特征信息,减轻因优化主干网络带来计算量增加的负担;最后通过改进 CIOU 损失,降低预测框与真实框之间的误差值,有效的解决因目标过小或者有遮挡时模型出现漏检误检等问题;在公开的 PASCAL VOC 数据集上进行验证,结果表明:RFAL YOLOv4 模型的 mAP 达到了 79.5%,比原模型提升了 5.5%,改进后的模型具有较高的鲁棒性。

关键词:目标检测;YOLOv4;Res2Net;注意力机制;CIOU

Detection and Algorithm for the YOLOv4 Object Based on Res2Net Fusion Attention Mechanism

ZHANG Xiang, LIU Zhenkai, YE Na, ZHAO Yanzhen
(School of Information and Control Engineering, Xi'an University of Architecture and Technology, Xi'an 710311, China)

Abstract: Aiming at the problems of missed detection, false detection and difficult detection with occlusions in traditional object detection algorithm, A Res2Net fusing with attention learning YOLOv4 (Res2Net fusing with attention learning YOLOv4, RFAL YOLOv4) object detection model is proposed. Firstly, to increase the receptive field of the model and obtain more semantic information of the feature map, by constructing a hierarchical class residual connection in a residual block, the Res2Net is introduced to replace the ResNet residual network structure in the original YOLOv4 backbone network, the model can obtain the finer features, at same time, the receptive field of the model is increased. Secondly, the attention mechanism is introduced to obtain the key feature information, and the residual network is integrated with the attention mechanism to reduce the burden of increased computation caused by optimizing the backbone network. Finally, the CIOU loss is improved to reduce the error between the prediction box and the real box, and the problem of missed or false detection with occlusions is effectively solved. The public PASCAL VOC data set is used to verify the improved model. The results show that the mAP of the RFAL YOLOv4 model reaches 79.5%, which is 5.5% higher than that of the original model. The improved model has better robustness.

Keywords: object detection; YOLOv4; Res2Net; attention mechanism; CIOU

0 引言

目标检测作为计算机视觉的重要研究领域,是解决图像描述、场景理解、语义分割、目标追踪和事件检测等更高层次视觉任务的基础。随着深度学习被应用于计算机视觉研究中,基于卷积神经网络的目标检测算法成为目标检测算法的主流。

基于深度学习的目标检测算法分为两阶段的目标检测算法和单阶段的目标检测算法。由学者 Girshick 提出的 R-CNN 算法[1]就是两阶段目标检测算法的代表,该算法将人工干预用 Selective Search[2]方法进行替换,利用卷积神经网络(CNN,convolutional neural networks)提取特征。使用支持向量机(SVM,support vector machine)对获取的特征进行分类。通过 PASCAL VOC 数据集[3]对 R-CNN 检测算法进行验证,准确率为 58.5%。因 R-CNN 在全连接层需要对图像进行裁剪、压缩为固定大小,最终影响检测结果的准确性。于是 Kaiming He 等人提出了多尺度空间金字塔网络[4](spatial pyramid pooling,SPP-net),SPP-net 的引

收稿日期:2022-02-28; **修回日期**:2022-04-18。
基金项目:陕西省自然科学科学基础研究计划资助项目(2018JM6080)。
作者简介:张 翔(1972-),男,陕西咸阳人,博士,副教授,硕士生导师,主要从事机器学习,深度学习方向的研究。
叶 娜(1979-),女,陕西西安人,博士,副教授,硕士生导师,主要从事机器学习、增强现实和智能信息处理方向的研究。
引用格式:张 翔,刘振凯,叶 娜,等.Res2Net 融合注意力机制的 YOLOv4 目标检测算[J].计算机测量与控制,2022,30(9):213-220,227.

入有效解决 R-CNN 对图像处理时导致的不完整性问题。由于 SPP-net 的特征提取是基于多尺度空间金字塔池化，需要花费大量时间在训练以及检测上，因此 Fast R-CNN[5] 将 SPP-net 池化替换为感兴趣区域池化（region of interest pooling，ROI pooling），节省检测时间和训练时间，并将 SVM 替换为 softmax，节省存储空间。Faster R-CNN[6] 对候选区的选择方式进行了改进，使用区域建议网络[6]（RPN，region proposal network）。将 RPN 嵌入到 Faster R-CNN 中去，实现真正意义上端到端的检测与训练。

2016 年 Redmon[7] 等人改变以往的两阶段检测思路，采用单阶段检测，这种全新的检测方法取名为 YOLO（You Only Look Once），其网络架构受到 GoogleNet[8] 图像分类的启发，将目标检测视为回归问题，看成一个单一神经网络，利用整张图像预测所有类的边界框。模型将图像分为网格，每个网格负责预测对象边界框和边框置信度，置信度反映了模型预测的准确性。

因 YOLO 需要对置信度位置进行推测，产生大量的 Prediction box，增加计算量影响检测速度。YOLOv2[9] 借鉴 Faster R-CNN 中 anchor box 思想，利用锚框对目标位置进行估计，减少 Prediction box 的数量，提高了检测速度。但 YOLOv2 对小目标检测精度不高，Joseph Redmon 提出 YOLOv3[10] 并引入了新的思想，主干网络采用 DarkNet-53，借鉴残差网络算法，增加检测模型的深度。YOLOv3 受特征金字塔网络[11]（FPN，feature pyrimad networks）启发，融合多个尺度特征网络，提升模型检测精度。Alexey Bochkovskiy 等人[12] 提出了 YOLOv4，模型检测精度以及检测速度相比于其他的 YOLO 系列得到提升，Liu 等人提出了 SSD[13] 模型，使用单个独立的过滤器进行检测，每个过滤器都进行图像特征提取，并将获取的特征传送到检测器中，以便执行对尺度检测，经过改进 SSD 的检测速度要比 YOLO 更加快速，而且检测的精准度相比于 Faster R-CNN 还要精确。

随着单阶段目标检测效率的提升，将该检测方式应用到其它检测领域已经成为热点问题。张兴旺团队将 Tiny-YOLOv3 应用在无人机地面目标跟踪上[14]，通过应用卡尔曼滤波器实现目标的有效跟踪。随着 2020 年疫情变得严峻，将目标检测与疫情防控融合，实现疫情的有效控制，于硕团队将 YOLOv5 应用到口罩佩戴检测中[15]，面对复杂的检测环境，可以有效进行防控工作。

虽然 YOLOv4 提升目标检测的精度，但对于细微层次上的特征表达不那么敏感，当检测的目标过小或者有遮挡时，模型会出现漏检以及误检等问题。

针对 YOLOv4 模型中存在的问题，提出 RFAL YOLOv4 模型，具体改进为引入 Res2Net，替换原有 ResNet，实现细小特征获取；利用注意力机制获取关键信息，减轻模型计算负担；改进 CIOU 损失函数，解决对有遮挡目标检测出现的漏检问题，提升模型的检测准确率。

1 YOLOv4 算法介绍

传统 YOLOv4 是以 CSPDarkNet53[16] 为主干网络，先将输入为 3 通道的图像处理为 32 通道，然后再输入到主干网络中进行卷积操作。基础主干网包含了 72 个卷积层，对输入的图像进行浅层特征提取，经过多域金字塔池化层（SPP）以及路径聚合网络[17]（PANet），将提取的浅层特征进行深层处理，最终将获取到的特征输入到检测头，输出 3 个大小不同的 heatmap。具体的 YOLOv4 算法结构如图 1 所示。

YOLOv4 输出为 3 个大小不同的 heatmap，对不同的目标进行检测。由于本文输入模型的图像像素值大小为 416 * 416，因此经过卷积操作以及池化操作后，输出的检测头像素值分别是 52 * 52、26 * 26、13 * 13。虽然 YOLOv4 在以前的 YOLO 上做了极大的优化，但还会出现漏检以及误检的目标，影响模型的检测准确率。

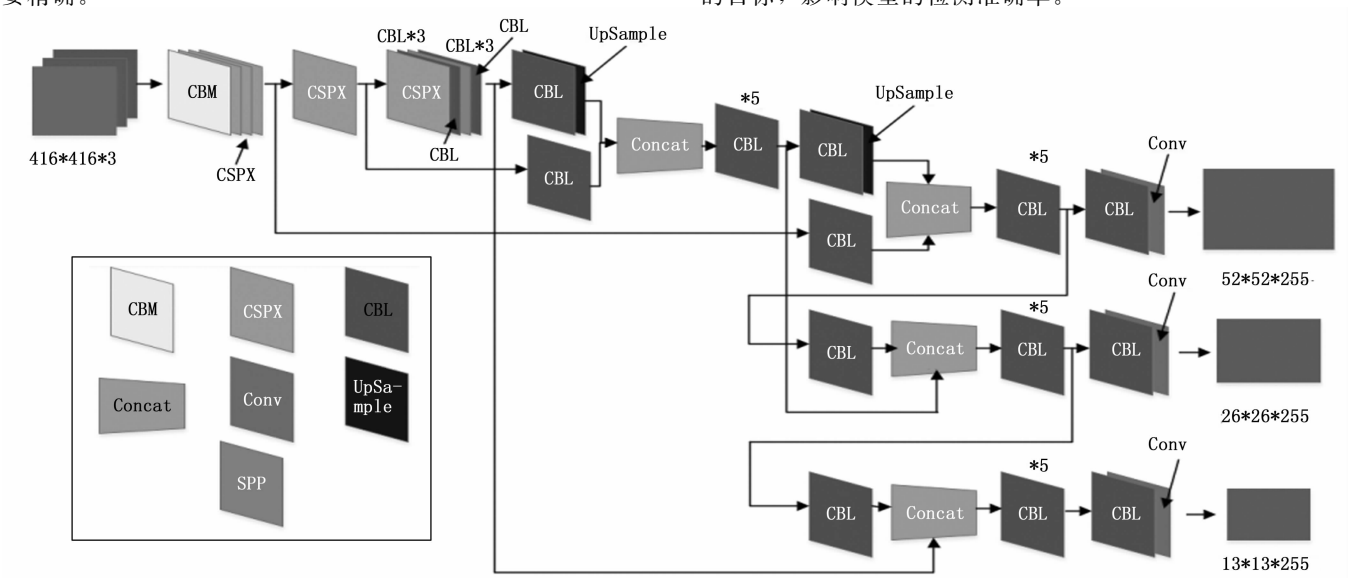

图 1 YOLOv4 算法结构图

2　RFAL YOLOv4

为改善 YOLOv4 算法在复杂背景下检测小目标困难，RFAL YOLOv4 首先在原始网络中改进 ResNet 残差网络，并融合注意力机制，实现卷积网络对图像特征细微获取，其次改进原有的 CIOU 损失函数，可以实现遮挡目标的有效识别。原始 YOLOv4 网络在 SCPResNet 结构的 part2 部分进行残差网络循环操作，RFAL YOLOv4 网络不仅在 CSPRes2Net 结构的 part2 部分进行组内残差循环操作，并且在完整的循环操作后融合注意力模块，网络结构对比如图 2 所示。

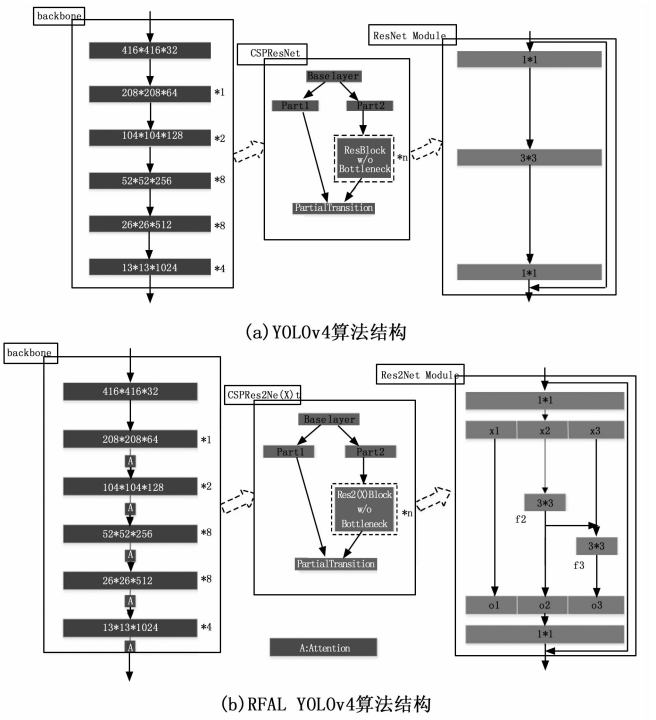

图 2　YOLOv4 与 RFAL YOLOv4 骨干网络对比

2.1　引入 Res2Net 网络结构

Res2Net[18] 在传统的 ResNet[19] 上进行改进，模型获取更细微图像特征。它通过在一个残差块内构造层次化的残差连接，增加了每个网络层的接受域。

ResNet 只是单一的残差操作，它的思想是将原有网络特征利用卷积操作一分为二，这两个结构块输入特征相同，一块不进行卷积操作，另一块进行卷积操作，然后将没有进行卷积操作的那一块特征与参加卷积操作的输出特征进行合并。

Res2Net 网络结构先进行分组，在进行残差操作，它的思想是首先将输入特征分组，其中一组过滤器先进行卷积操作，提取输入信息的特征，然后将获取到的特征和另外一组发送过来准备输入的特征一起输入到下一个过滤器，对上述过程不断的重复执行，它的目的是对输入特征进行完全处理，处理完之后该操作结束，最后连接这些特征图，并将连接好的特征图传递到一个 1 * 1 的过滤器，融合所有特征。在特征传递过程中，输入特征可以以任何路径进行转化，当通过和上一次卷积相同的 3 * 3 过滤器时，由于卷积操作的原因使得感受野增加。网络结构对比如图 3 所示。

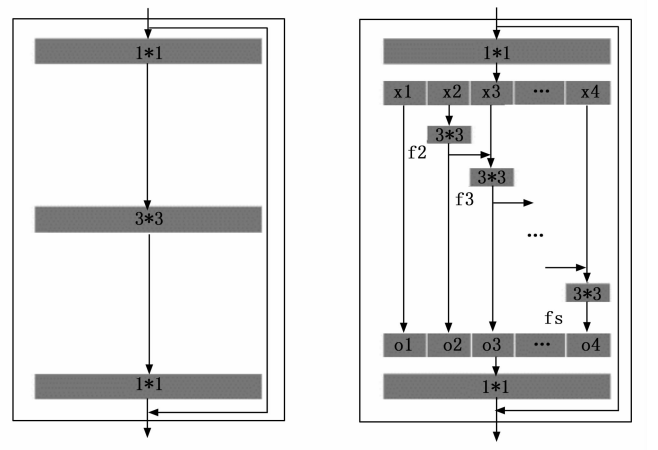

图 3　网络结构对比图

引入 Res2Net 的网络结构如图 2 所示。首先将 416 * 416 像素的图像输入到 CSPDarknet53 网络结构中，经过标准化、归一化、激活处理之后，输入到一个残差块网络结构中，然后将输入特征分为两个部分对应图 2（b）中的 part1 和 part2。part1 进行基础卷积即标准化、归一化、激活。part2 对输入的特征进行循环残差，此时的残差块为 Res2Net 网络结构。图像特征经过 1 * 1 卷积之后，均匀的分割为大小相同的 3 个特征子集，将他们记为 x_i，其中 $i\in\{1,2,3\}$。分割的每一个特征层 x_i 与输入的特征层的特征空间大小相同，但是通道数变为原来的 $\frac{1}{3}$。并且除了 x_1，其它的 x_i 都有一个对应的 3 * 3 卷积，由于增加了模型的通道数，并且进行卷积操作时，可以获取更多的特征，每个卷积层用 f_i 表示，使用 o_i 表示对应的卷积层 f_i 的输出。融合 x_i 与 f_{i-1}，通过一个 3 * 3 卷积，最终输出。o_i 可以写成公式（1）：

$$o_i=\begin{cases}x_i\\ f_i(x_i)\\ f_i(x_i+o_{i-1})\end{cases}\tag{1}$$

虽然引入组内残差结构，增加了网络的卷积操作，可以有效的获取图像上下文特征，但增加了模型的复杂度。相比于原始的 CSPDarknet 网络，改进后的网络每经过一次 CSP 网络，都要进行特征分割并进行卷积操作，增加了模型的计算量，影响了模型的检测实时性，但模型的检测精度得到提升。

2.2　引入注意力机制

由于组内残差对特征过度细小化提取，导致模型计算量增加。因此引入注意力机制获取关键信息降低 Res2Net 计算量。

2018 年提出（CBAM，convolutional block attention module）[20]，一种既考虑不同通道像素相关性，又考量空间

像素影响的注意力算法。并且进行组内残差连接之后，虽然可以获取到更深、更有判别力的特征图信息，但这也会造成特征图空间信息出现丢失。

CBAM 首先进行通道域处理，将输入的特征进行全局最大池化和全局平均池化，然后送入一个两层的神经网络并进行激活操作，最后将这两层网络进行加和在经过 sigmoid 激活，生成 channel attention feature，将得到的通道注意力特征和输入特征进行乘法操作，最终生成 spatial attention 模块需要的输入特征；其次进行空间域处理，将 channel attention 模块输出的特征做一个基于 channel 的全局最大池化和全局平均池化，得到两个 H * W * 1 的特征图，然后将这两个特征图做拼接操作，然后进行 7 * 7 卷积，进行降维操作，在经过 sigmoid 生成 spatial attention feature，最后将 channel attention 输出的特征和该模块处理的特征进行乘法操作，得到最终的 Refined Feature。因此引入 CBAM 注意力机制模块可以有效地避免信息丢失等问题。CBAM 注意力模块如图 4 所示。

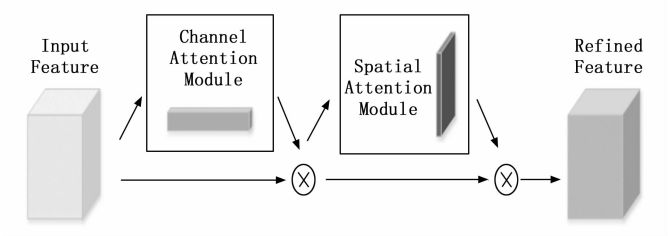

图 4 CBAM 网络结构图

将 CBAM 注意力模块融合到组内残差网络中，模型在进行训练时，图像经过主干网络进行特征提取，首先进行归一化、前向传播以及激活操作，然后利用卷积操作将处理后的图像特征分为两个部分，这两部分特征相同，其中一部分特征保留，而另一部分特征输入到组内残差卷积块中，模型对图像特征进行卷积操作，并获取细小特征，利用卷积操作将这两个部分的特征合并，最后将提取的特征输入到 CBAM 注意力模块，利用注意力模块对获取特征进行关键信息提取，使用该注意力模块增加模型感受野，从而使模型具有更好的检测效果。以及引入 CBAM 注意力机制，可以有效抑制背景信息干扰。融合后的网络结构图如图 5 所示。

替换 ResNet 为 Res2Net 并在组内残差网络之后融入 CBAM 注意力模块，在不同的像素下输出的特征图像如图 6 所示，上半部分为原始 YOLOv4 处理后的图像特征输出，下半部分为 RFAL YOLOv4 特征输出，RFAL YOLOv4 训练的图像像素输入值设置为 416 * 416，经过卷积之后网络的三层输出像素分别是 52 * 52、26 * 26、13 * 13。

通过 Res2Net 和 CBAM 融合后输出特征图对比发现，引入注意力机制，52 * 52、26 * 26 像素输出提高目标与背景的区别，替换 ResNet 网络之后，RFAL YOLOv4 可以获取更多的特征，尤其是在 52 * 52 像素特征输出时最为明显。

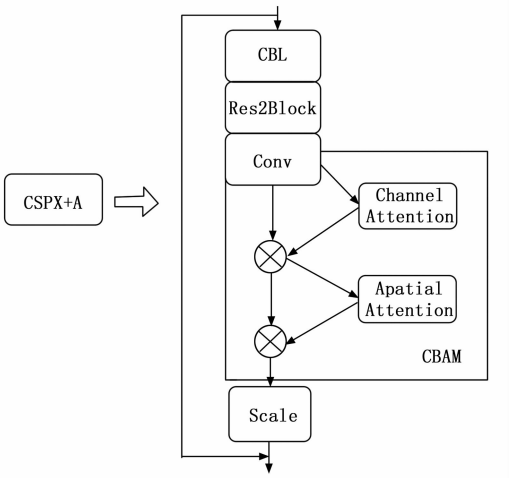

图 5 CBAM-Res2Net 网络结构图

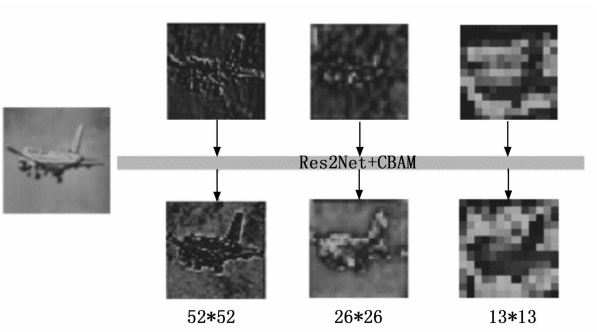

图 6 特征输出对比图像

2.3 改进 CIOU 损失

目标检测任务中，利用损失函数反应 Real Box 和 Prediction Box 的误差，其中 IOU[21] 损失函数的范围在（0，1）之间，具有尺度不变性，IOU 损失可以评价各种形状的匹配程度。所以将 IOU 损失引入到 YOLO 检测中。IOU 损失如公式（2）所示。

$$IOU = \frac{A}{B} \tag{2}$$

式中，A 表示 Prediction Box 与 Real Box 的交集即获取的交集区域的面积，B 表示 Real Box 与 Prediction Box 的并集即获取的并集部分的面积。利用获取的交集的面积和并集的面积作比，最终得到 IOU。

但当 Real Box 与 Prediction Box 不相交时 IOU 为零，不能预测 Real Box 与 Prediction Box 之间的误差值；当 Prediction Box 与 Real Box 交集、并集相同，此时得到的 IOU 相同，但位置不一致时，模型获取到的图像特征是不相同的，因此 IOU 损失函数不满足此时的需求。

因此 GIOU[22] 损失在 IOU 损失的基础上进行优化，GIOU 损失如公式（3）所示。

$$GIOU = IOU - \frac{|C - (A \cup B)|}{|C|} \tag{3}$$

式中，$\boldsymbol{C}$ 表示 Real Box 与 Prediction Box 的最小外接矩形，但当 Prediction Box 位于 Real Box 内部时，GIOU 会退化为 IOU 损失，无法优化相对位置，收敛缓慢。由于出现损失退化的问题，所以对损失函数进行优化，产生 DIOU 损失[23]。

DIOU 损失即计算 Prediction Box 与 Real Box 中心之间的距离 d_c^2，获取最小外接矩形对角线距离 d_s^2，DIOU 损失如公式（4）所示。

$$DIOU = IOU - \frac{d_c^2}{d_s^2} \tag{4}$$

DIOU 保留了 GIOU 的优点，对两个框的中心点进行度量，使其进行快速收敛，并且当两个框的中心点重合时，DIOU 才退化为 IOU，虽然 DIOU 在前人的基础上做了改进，但该损失有个缺陷是当 Prediction Box 与 Real Box 距离一致但是框的形状不一致，其计算的 DIOU 结果相同，无法进行模型的优化。所以对其进行优化，产生 CIOU 损失。

CIOU 损失考虑到同一中心点位置的 Prediction box 的形状不同，在 DIOU 损失函数的基础上对其添加惩罚项。用于 Prediction box 形状的区分。

CIOU 损失如公式（5）所示。

$$CIOU = IOU - \frac{d_c^2}{d_s^2} - \alpha v \tag{5}$$

$$\alpha = \frac{v}{(1 - IOU) + v} \tag{6}$$

$$v = \frac{4}{\pi^2}(\arctan\frac{w^g t}{h^g t} - \arctan\frac{w}{h})^2 \tag{7}$$

式中，α 用以平衡参数，v 表示宽高比的倾斜角度。

YOLOv4 模型训练时，当出现划定的多个位置不同的 Prediction Box 形状相同，且它们到 Real Box 中心点距离相同，计算得到多个 Prediction Box 的 CIOU 损失值一致，导致模型在回归操作时会舍弃相同的 Prediction Box，因此出现漏掉图像部分特征的问题。当模型进行检测任务时，如果检测目标恰好被某个物体遮挡，遮挡部分的 Prediction box 与没有被遮挡部分的 Prediction box 进行损失计算时，它们的损失函数值一致，这就导致目标在被遮挡时出现检测误差。

对上述问题，RFAL YOLOv4 在 CIOU 的基础上添加了惩罚项，计算 Prediction box 左上角到 Real box 中心距，以此区别于 Prediction box 与 Real box 中心距相同、Prediction box 形状相同、位置不同时的特殊情况。损失函数计算公式如（8）所示：

$$Loss = 1 - Loss_{YOLOv4} - \beta * t \tag{8}$$

β 计算公式如（9）所示：

$$\beta = \frac{t}{(1 - IOU) + t} \tag{9}$$

t 的计算公式如（10）所示：

$$t = I(a,b) \tag{10}$$

式中，a 表示 Prediction box 到 Real box 左上角的距离，b 表示 Prediction box 到 Real box 中心点的距离，I 表示坐标距离的比值，β 表示平衡 t 所添加的参数项。

2.4　模型训练

采用的数据集为官方的 VOC 数据集，首先运行 voc2yolo4. py 文件，将数据集划分为训练集、测试集、验证集，然后设置 classes 为 aeroplane、bicycle、bird 等 20 个类别，然后设置模型输入图像的像素，输入 shape 大小为 416 * 416，设置 anchor _ num 为 9，进行聚类操作生成九个聚类框的坐标，参数设置完成之后进行模型训练。

2.4.1　超参数设置

1）修改 VOC _ label. py 文件中的类别标签，将其修改为要识别的目标类别。由于本文采用 VOC2007＋2012 数据集，进行模型训练时，在该文件下将 classes 分别修改为对应的分类类别。

2）修改 obj. data 文件中的 classes、修改为训练数据集的类别数，因此设定 classes 分别为 20。设定训练数据集路径、模型训练时生成的权重值存储文件路径，cfg/obj. names 文件为模型训练时各类别名称。

3）修改 obj. names 文件中进行分类的类别，进行 VOC 数据训练时，在该文件中设置类别为 aeroplane、bicycle、bird 等类别。

4）YOLOv4. cfg 文件：此文件主要为模型的网络结构，文件开头部分描述了用于训练网络的一些超参数，剩余部分为图像特征提取的主干网络。修改一部分超参数使其适应本文数据集和训练环境。文件中将 batch 设置为 64、subdivisions 设置为 16（主要取决于模型训练的 GPU 内存大小）。max _ batches＝40 000（迭代次数为训练类型 classes * 2 000）、steps＝32 000，36 000（设置为 max _ batches 的 80%、90%）、classes＝20，anchors 为 kmeans 聚类之后获取的先验框坐标、［yolo］－layers 下的三处 $filters$＝75。$filters$ 的计算公式如式（11）所示：

$$filters = (5 + classes) \times 3 \tag{11}$$

式中，$filters$ 表示卷积输出，$classes$ 表示类别数。

2.4.2　预编译

首先运行 test. py 文件，将数据集划分为训练集、测试集以及验证集，然后运行 VOC _ label. py 文件，获取 Annotations 标签文件中的标注文件属性。最后运行 kmeans _ for _ anchors. py 文件，使用 kmeans 算法，对模型训练时的数据集进行聚类操作，并生成聚类框的坐标值。

3　实验结果与分析

3.1　数据集

本次实验采用的数据集来自 PASCAL VOC 2007[4] 和 PASCAL VOC 2012，其中包含有人、猫、狗、马、自行车、电视、沙发、鸟等等 20 个种类 22 077 张图片。本次实验将数据集随机划分为验证集、训练集、测试集，训练集有 19 870 张图片，测试集有 1 985 张，验证集有 222 张，使用 19 870 张不同种类的图片实验模型的预训练，使用 1 985 张图片进行模型的测试。实验数据如表 1 所示。

表 1 实验数据划分

	VOC2007	VOC 2012	合计
训练集	4 457	15 413	19 870
测试集	445	1 540	1 985
验证集	50	172	222
合计	4 952	17 125	22 077

3.2 实验环境

3.2.1 实验硬件环境

由于该实验需要对大量的图数据进行处理，所以需要利用 GPU 进行计算。表 2 为实验所需要的硬件环境。

表 2 硬件环境

名称	型号	数量
CPU	Intel(R) Core(TM) i7－9700F	1
主板	技嘉(GIGABYTE)B450 AORUS PRO	1
GPU	RTX 2080Ti(显存:11G)	1
内存	海盗船 DDR4 3200 8G	2
固态硬盘	三星(SAMSUNG)500G	1
机械硬盘	西数数据 2T	1

3.2.2 实验软件环境

本实验所使用的模型为 RFAL YOLOv4，部分实验采用的权重参数如表 3 所示。

表 3 软件环境参数值表

参数名称	参数值
Cuda	cuda_10.0.130_411.31_win10
Cudnn	cudnn-10.0-windows10-x64-v7.4.1.5
Python	3.6.8

3.3 模型性能比较

采用 AP 和 mAP 进行模型性能评价，将本文改进的检测模型 RFAL YOLOv4 和 Fast、Faster、SSD、YOLO、YOLOv2-v4 目标检测模型进行对比，各个类别的检测结果如表 4 所示，RFAL YOLOv4 在检测精度上得到相应的提升。通过图表对比各模型检测结果，对某个类别检测精度最高项，采用斜体加粗的方式进行标记。余丽仙团队改进 SSD[24] 六个检测类别精度占优，YOLOv2 两个检测类别精度占优，YOLOv3 一个检测类别精度占优，YOLOv4 三个检测精度占优，RFAL YOLOv4 有九项检测精度占优，其中 YOLOv4 与 RFAL YOLOv4 有一项并列占优。RFAL YOLOv4 模型相比于传统 YOLOv4 模型，mAP 提升了 5.5%，相比于改进 SSD 目标检测算法，mAP 仅差 0.1%，但 RFAL YOLOv4 在具体类别检测中占相对高的优势，相比于 YOLOv2 提升了 6.1%，相比于 YOLOv3 提升了 5.6%。通过图 8 可以直观的看出，RFAL YOLOv4 模型基本处于最高点。

网络模型的评价指标表现为两方面，分别为检测模型的精度和速度，RFAL YOLOv4 相比于传统 YOLOv4 模型在检测的速度上有所下降，但是模型的检测精度提高了 5.5%。具体的模型对比如表 5 所示。

YOLOv4 和 RFAL YOLOv4 在训练时检测精准度对比如图 7 所示，其中横坐标表示模型在各个类别上的检测精准度，纵坐标表示模型的类别，mAP 表示检测模型的最终检测精准度。

3.4 检测结果对比

通过检测结果图更加直观的展示 RFAL YOLOv4 有效性。图 9 为 YOLOv4 与改进 RFAL YOLOv4 在测试集上的检测结果。采用 PASCAL VOC 2007＋2012 train 训练集。图像大小的输入均为 416 * 416 像素值。

$a1-k1$ 为传统 YOLOv4 算法的检测结果，图 a2－k2 为 RFAL YOLOv4 算法的检测结果。通过 $a1$、$a2$、$b1$、$b2$ 对

表 4 VOC 测试数据集下 20 类检测精度对比

Method	data	mAP/%	aero	bike	bird	boat	bottle	bus	car	Cat	chair	cow
Fast[5]	07＋12	70.0	82.3	78.4	70.8	52.3	38.7	77.8	71.6	89.3	44.2	73.0
Faster[6]	07＋12	73.2	84.9	79.8	74.3	53.9	49.8	77.5	75.9	88.5	45.6	77.1
YOLO [7]	07＋12	63.4	77.0	67.2	57.7	38.3	22.7	68.3	55.9	81.4	36.2	60.8
SSD[24]	07＋12	79.6	78.3	***88.7***	77.1	***73.6***	55.8	87.2	***87.1***	90.1	62.5	85.3
YOLOv2[9]	07＋12	73.4	86.3	82.0	74.8	59.2	51.8	79.8	76.5	90.6	52.1	78.2
YOLOv3[11]	07＋12	73.9	89.0	68.0	***83.0***	64.0	56.0	86.0	80.0	90.0	57.0	78.0
YOLOv4[12]	07＋12	74.0	***90.0***	66.0	73.0	57.0	64.0	***88.0***	83.0	77.0	66.0	87.0
RFAL YOLOv4	07＋12	79.5	84.0	82.0	80.0	59.0	***83.0***	***88.0***	87.0	87.0	60.0	***95.0***
Method	data	mAP/%	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv
Fast[5]	07＋12	70.0	55.0	87.5	80.5	80.8	72.0	35.1	68.3	65.7	80.4	64.2
Faster[6]	07＋12	73.2	55.3	86.9	81.7	80.9	79.6	40.1	72.6	60.9	81.2	61.5
YOLO[7]	07＋12	63.4	48.5	77.2	72.3	71.3	63.5	28.9	52.2	54.8	73.9	50.8
SSD[24]	07＋12	79.6	***78.7***	87.6	90.1	87.6	82.8	***51.3***	80.7	79.8	89.4	***78.4***
YOLOv2[9]	07＋12	73.4	58.5	***89.3***	82.5	83.4	81.3	49.1	77.2	62.4	83.8	68.7
YOLOv3[11]	07＋12	73.9	62.0	81.0	66.0	61.0	83.0	41.0	78.0	78.0	86.0	67.0
YOLOv4[12]	07＋12	74.0	59.0	86.0	77.0	82.0	85.0	44.0	78.0	63.0	86.0	70.0
RFAL YOLOv4	07＋12	79.5	68.0	81.0	***95.0***	***96.0***	***89.0***	33.0	***84.0***	***81.0***	***92.0***	68.0

表 5　VOC 数据集下模型检测速度对比

检测模型	数据集(VOC)	mAP(%)	检测速度(FPS)
Fast R-CNN[5]	2007+2012	70.0	0.5
Faster R-CNN VGG[6]	2007+2012	73.2	7.0
Faster R-CNN ResNet[6]	2007+2012	76.4	5.0
SSD[24]	2007+2012	79.6	37.3
YOLO[7]	2007+2012	63.4	45.0
YOLOv2[9]	2007+2012	73.4	43.1
YOLOv3[11]	2007+2012	73.9	78.0
YOLOv4[12]	2007+2012	74.0	96.0
RFAL YOLOv4	2007+2012	79.5	84.0

比，原始 YOLOv4 模型出现误检，将没有出现的目标误检为椅子，将目标中的马和人误检，RFAL YOLOv4 有效的识别出目标中的检测对象。$c1$、$c2$、$d1$、$d2$ 对比，YOLOv4 算法出现漏检，未能识别桌子以及桌子上的瓶子，$d1$ 图像中漏检了椅子，而 RFAL YOLOv4 算法对桌子、椅子、瓶子识别率高。通过 $e1$、$e2$ 对比，在检测的对象较多时，YOLOv4 出现漏检，未能充分检测图像中的鸟类，RFAL YOLOv4 算法精准的识别出图像中的鸟。

为验证 RFAL YOLOv4 算法对小目标检测效果，从 VOC 数据集中挑选出小目标图像进行检测，图中 $f1-h1$ 为传统 YOLOv4 检测结果，$f2-h2$ 为 RFAL YOLOv4 算法检测结果，通过 $f1$、$f2$ 对比，RFAL YOLOv4 精准的识别出图像中的船舶，通过 $g1$、$g2$ 对比，RFAL YOLOv4 算法对汽车识别效果好，通过 $h1$、$h2$ 对比，RFAL YOLOv4 算法可以精确识别图像中出现的羊。

通过在 VOC 数据集中选取 200 张有遮挡图像，组成小型遮挡数据集进行检测，通过实验最终得到改进 CIOU 损失，如表 6 所示，模型的检测精度相比于原始 YOLOv4 检测模型提升了 3.1%，相比于文献 21 中 SSD 算法 mAP 高 0.5%。

表 6　遮挡目标检测结果对比

模型	mAP(%)	FPS
YOLOv4[12]	75.2	92
SSD[24]	77.8	37.3
RFAL YOLOv4	78.3	84

为了验证在有遮挡的情况下 RFAL YOLOv4 算法的鲁棒性。通过实验验证 RFAL YOLOv4 算法对遮挡目标检测效果比较好。通过 $i1$、$i2$ 行人被遮挡后，RFAL YOLOv4 算法可以精准检测，通过 $j1$、$j2$ 行船几乎被完全遮挡时，RFAL YOLOv4 算法可以准确检测行船，通过对比 k1、k2，RFAL YOLOv4 对狗有好的识别能力。

通过检测结果直观的展示 RFAL YOLOv4 检测模型的有效性。检测图像

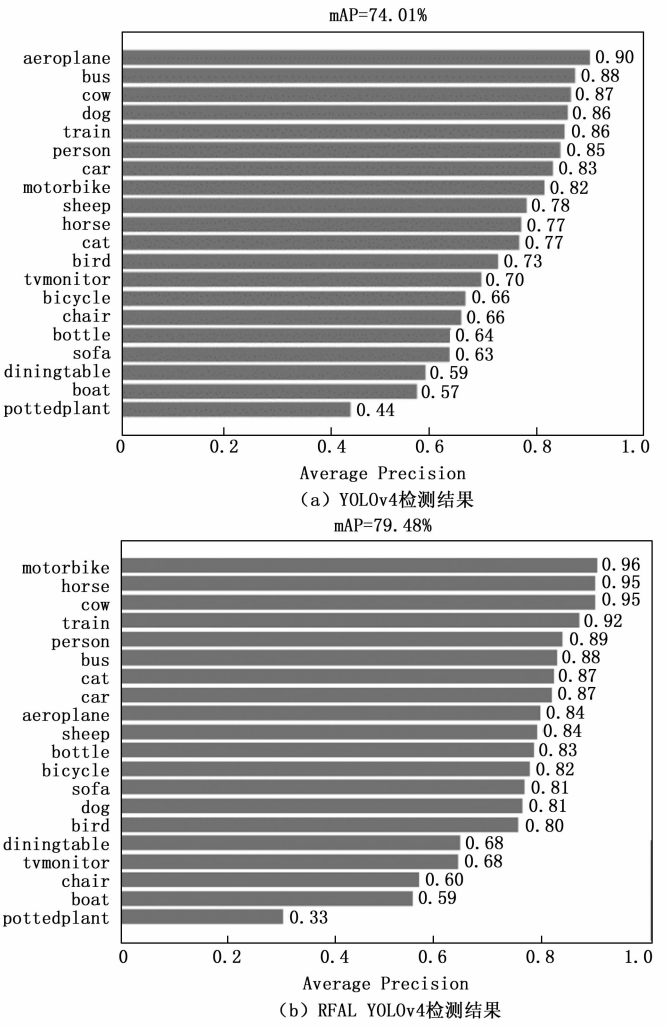

图 7　模型检测结果对比

中存在较多的检测目标，可以更多更精确的检测；检测对象中存在有较小的目标，RFAL YOLOv4 有更好的识别能力；检测对象中存在有遮挡，RFAL YOLOv4 模型也表现出更好的检测效果。

通过上述实验，充分的说明了 RFAL YOLOv4 模型有效的改善了漏检误检等问题。

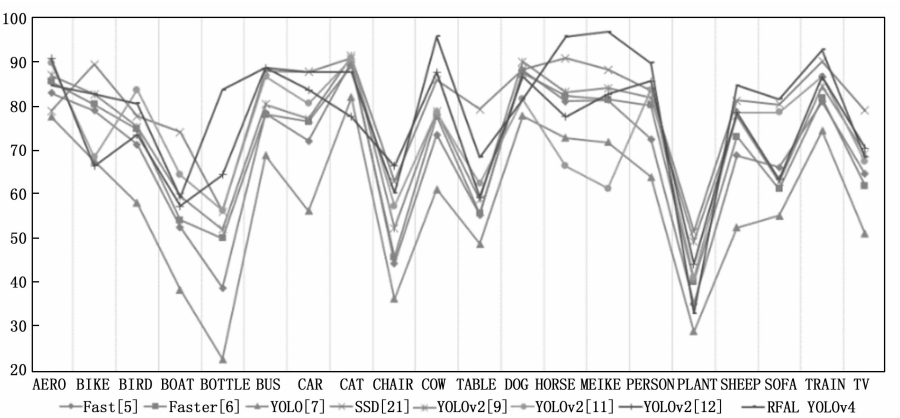

图 8　VOC 测试数据集 20 类别检测精度对比

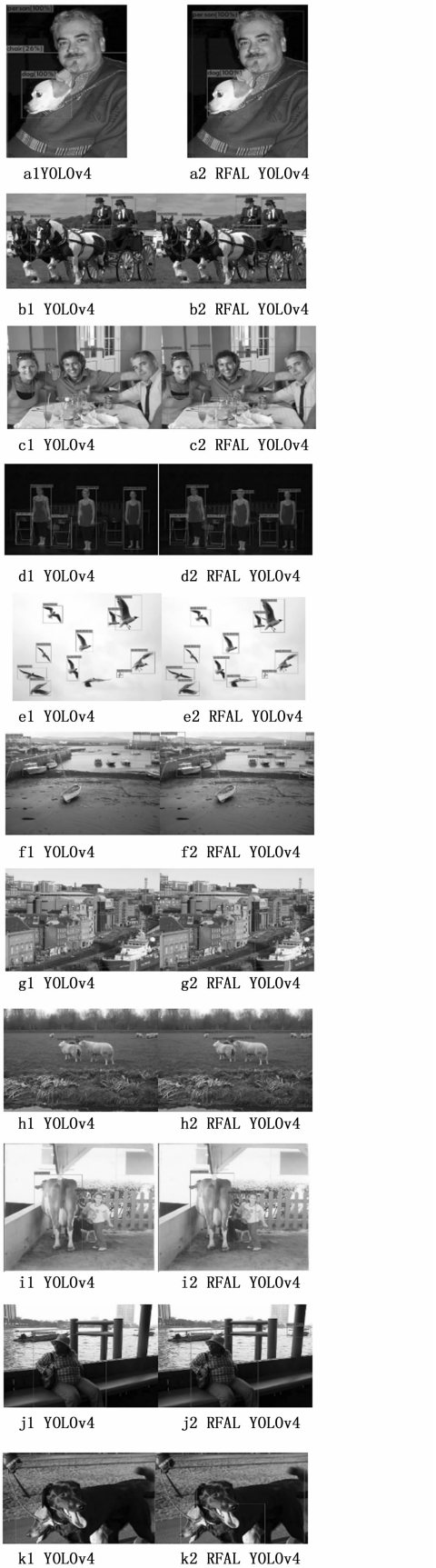

图 9 YOLOv4 与 RFAL YOLOv4 检测结果对比

4 结束语

考虑到传统的 YOLOv4 模型在检测小目标、有遮挡目标、以及在检测相似性目标时，识别能力较弱的情况。对传统模型进行优化，首先将传统 YOLOv4 网络中的 ResNet 网络更改为 Res2Net，可以获取细小的特征。为了不影响检测模型的实时性，将组内残差网络的特征子集分为 3 个。其次在主干网络后融入注意力机制，有效的区分目标与背景信息，从而增加了模型的识别能力。最后对模型的损失进行改进，有效的检测遮挡目标。使用公开的 VOC 数据集对模型进行对比验证，在测试集上模型的检测精度达到了 79.5%，在传统的检测模型上提高了 5.5%，充分的证明 RFAL YOLOv4 模型具有更好的鲁棒性。

参考文献：

[1] GIRSHICK R，DONAHUE J，DARRELL T，et al. Rich feature hierarchies for accurate object detection and semantic segmentation [C] //Proceedings of the 2014 IEEE Conference on Computer Vision and PatternRecognition，2014：580－587.

[2] UIJLINGS J R R，VAN DE SANDE KEA，GEVERS T，et al. Selective search For object recognition [J]. International Journal of Computer Vision，2013，104 (2)：154－171.

[3] EVERINGHAM M，WINN J. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Development Kit [J]. International Journal of Computer Vision，2006，111 (1)：98－100.

[4] HE K，ZHANG X，REN S，et al. Spatial pyramid pooling in peep convolutional net-works for visual recognition [J]. IEEE Transa-ctionson Pattern Analysis&Machine Intelli-gence，2014，37 (9)：1904－1916.

[5] GIRSHICK R. Fast R-CNN [C] //Proceedings of the 2015 IEEE International Conference on Computer Vision. 2015：1440－1448.

[6] REN S，HE K，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，2016，39 (6)：1137－1149.

[7] REDMON J，DIVVALA S，et al. You only look once：unified，realtime object detection [C] // Nevada：IEEE Conference on Computer Vision and Pattern Recognition，2016：779－788.

[8] SZEGEDY C，LIU W，et al. Going deeper with convolutions [C] //Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. 2015：1－9.

[9] REDMON J，FARHADI A. YOLO9000：better，faster，stronger [C] //Hawaii：IEEE Conferenceon Computer Vision and Pattern Recognition，2017：7263－7271.

[10] REDMON J，FARHADI A. YOLOv3：anincremental improvement [J]. arXiv preprintarXiv：1804. 02767，2018：https://arxiv. xilesou. top/abs/1804. 02767.

[11] LIN，TSUNG-YI，et al. Feature pyramid networks for object detection [C] //proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)，2017：2117－2125.

(下转第 227 页)