第 59 卷 第 8 期/2022 年 4 月/激光与光电子学进展 研究论文

激光与光电子学进展 先进成像

基于 RDM-YOLOv3 的头部检测

刘竣文[1]，张永军[1*]，李智[1]，赵勇[2]，冉新宇[1]，崔忠伟[3]，牛梦佳[1]

[1]贵州省智能医学影像分析与精准诊断重点实验室，贵州大学计算机科学与技术学院，贵州 贵阳 550025；

[2]北京大学深圳研究生院信息工程学院，广东 深圳 518055；

[3]贵州师范学院大数据科学与智能工程研究院，贵州 贵阳 550018

摘要 现有的通用检测方法在小目标检测上仍存在漏检率较高的问题。为了提高头部的检测率，在 YOLOv3 基础上提出了 ResNet DenseNet MDC（Mixed Dilated Convolution）YOLOv3（RDM-YOLOv3）目标检测网络。首先改进了 YOLOv3 的特征提取网络 DarkNet-53，提出了一种基于 ResNet 和 DenseNet 的特征提取网络 RD-Net，以提取更多的语义信息。然后，使用不同膨胀率的空洞卷积对特征层进行采样，构建混合空洞卷积结构，提高对小目标的敏感度。使用 RDM-YOLOv3 与其他方法在 Brainwash 数据集和 HollywoodHeads 数据集上进行对比实验，AP（Average Precision）值分别达到了 93.1% 和 86.8%。所提方法的实验结果优于其他方法，对小目标的检测性能显著提升。

关键词 机器视觉；头部检测；小目标；卷积神经网络；特征提取网络；RDM-YOLOv3

中图分类号 TP391 **文献标志码** A **doi**：10.3788/LOP202259.0815011

Head Detection Based on RDM-YOLOv3

Liu Junwen[1], Zhang Yongjun[1*], Li Zhi[1], Zhao Yong[2], Ran Xinyu[1], Cui Zhongwei[3], Niu Mengjia[1]

[1]*Key Laboratory of Intelligent Medical Image Analysis and Precise Diagnosis of Guizhou Province, College of Computer Science and Technology, Guizhou University, Guiyang, Guizhou* 550025, *China*;

[2]*School of Information Engineering, Peking University Shenzhen Graduate School, Shenzhen, Guangdong* 518055, *China*;

[3]*Big Data Science and Intelligent Engineering Research Institute, Guizhou Education University, Guiyang, Guizhou* 550018, *China*

Abstract The existing general detection methods still have the problem of high missing rate in small target detection. To improve the detection rate of the head, the ResNet DenseNet MDC (Mixed Dilated Convolution) YOLOv3 (RDM-YOLOv3) target detection network is proposed on the basis of YOLOv3. Firstly, the feature extraction network DarkNet-53 of YOLOv3 is improved, and a feature extraction network RD-Net based on ResNet and DenseNet is proposed to extract more semantic information. Then, a mixed dilated convolution structure is constructed by sampling the feature layers using dilated convolution with different dilated rates to improve the sensitivity to small targets. Using RDM-YOLOv3 to compare with other methods on Brainwash dataset and HollywoodHeads dataset, the AP (Average Precision) values reached 93.1% and 86.8%, respectively. The experimental results are better than that of other methods, and the

收稿日期：2021-07-30；**修回日期**：2021-08-18；**录用日期**：2021-08-24

基金项目：国家自然科学基金（62062023）、贵州省教育厅创新群体研究项目（黔教合 KY 字[2021]022）

通信作者：*ljw778@126.com

performance of small target detection is significantly improved.

Key words machine vision; head detection; small targets; convolutional neural networks; feature extraction network; RDM-YOLOv3

1 引 言

行人检测是计算机视觉中的一个重要研究方向，可应用于人员识别[1]、行人跟踪[2]和自动驾驶[3]等领域。目前的检测方法在人群密集、遮挡严重场景中的行人检测效果差强人意，于是行人头部检测方法应运而生，监控视频中的人群计数[4]和头部检测是其重要的应用。随着卷积神经网络(CNN)和深度学习的发展，目标检测已经取得了长足进步[5]，但是在复杂人群场景中[6-7]，目标对象具有多样性、遮挡力强、动态模糊、分辨率低和稀有特征，头部检测仍然是一项艰巨的任务。

近年来，出现了许多基于深度学习的头部检测器[8-9]。大多数方法都是将头部检测视为一种特殊的目标检测。由于头部的尺度和外观各不相同，如何有效地提取特征来定位头部并将其和背景分离仍然是一个挑战。这些方法中表现比较好的是RCNN[10]头部检测器，它使用两种模型来进行头部检测：第一种是Global model，其给出多尺度的热力图来确定人头存在的概率[11]；第二种是Local model，其使用搜索建议(SS)来限制目标假设的集合。将两种model的线索组合在一起就是RCNN头部检测框架。Hariharan等[12]使用超列(Hyper column)，即将对应像素的网络所有节点的激活串联作为特征，进行目标的细粒度定位。SSD[13]和YOLO[14]采用了多尺度特征来检测物体，得出类别的概率和边界框坐标，其中YOLO在使用残差网络(ResNet)[15]进行特征提取的同时，使用空间特征金字塔(FPN)[16]完成上下文特征融合，这两个检测器的速度远快于Faster RCNN[17]。

尽管上述模型在图像的多个对象分类方面[18]取得了相当大的进步，但它们在检测小目标方面仍然存在不足，因为大多数模型使用最后一个卷积层的特征来进行对象检测。然而，最后的卷积层包含小对象的信息不足。由于头部检测问题中目标的占比很小，因此目前的方法不适合检测小目标。本文提出一种基于ResNet与DenseNet[19]的特征提取网络。RD-Net从特征重用的角度减少计算量、提高层间信息透过率，提取到更多语义信息；MDC模块对特征图用不同膨胀率的空洞卷积进行采样并进行融合，以扩大感受野、利用更多的细粒度特征信息，从而提高对小目标的敏感度。

2 YOLOv3原理

YOLOv3是一种单阶段算法，它将目标检测转换为一个回归问题。与Faster R-CNN相比，YOLOv3可直接获得目标位置和类别的预测信息，而不需要区域建议网络(RPN)[20]。YOLOv3网络将每个输入图像划分为$s\times s$个网格单元，网格以目标所在的真实框为中心，负责对其进行检测。针对每个网格单元定义B个边界框及其相应的置信度分数$P(C_i|O)$，其中C_i代表第i个类别，O代表目标。每个边界框都包含C类。如果目标的中心落在网格单元中，则$P(O)=1$；否则，$P(O)=0$。置信度得分定义为$P(O)\times R_{\mathrm{IOU}}$，它反映了网格单元包含目标的概率和边界框预测的精度，其中R_{IOU}为真实目标与检测目标的交并比。IOU表示边界框和真实框之间的重叠区域，可用A_{IOU}表示为

$$A_{\mathrm{IOU}}=\frac{A_{\mathrm{overlap}}}{A_{\mathrm{union}}},\tag{1}$$

式中：A_{overlap}为交集区域面积；A_{union}为并集区域面积。预测目标的具体类别分数可以表示为

$$P(C_i|O)\times P(O)\times R_{\mathrm{IOU}}=P(C_i)\times R_{\mathrm{IOU}}。\tag{2}$$

YOLOv3借鉴了空间特征金字塔网络的思想，该网络对每个输入图像进行5次下采样处理。对经特征提取的特征图进行下采样，得到输出特征图大小是输入图像的1/32。然后，YOLOv3将最后3个下采样特征图传送到3个不同尺度的检测层进行预测。3个尺度检测层的大小分别为13×13、26×26和52×52，分别负责检测大目标、中目标和小目标。深层特征图包含大量语义信息，而浅层特征图包含大量的细粒度信息。因此，为了进行特征融合，深层特征图通过上采样与浅层特征图级联。YOLOv3网络结构如图1所示。图1中RESi表示进行了i次残差堆叠，$i=1,2,\cdots,N$。

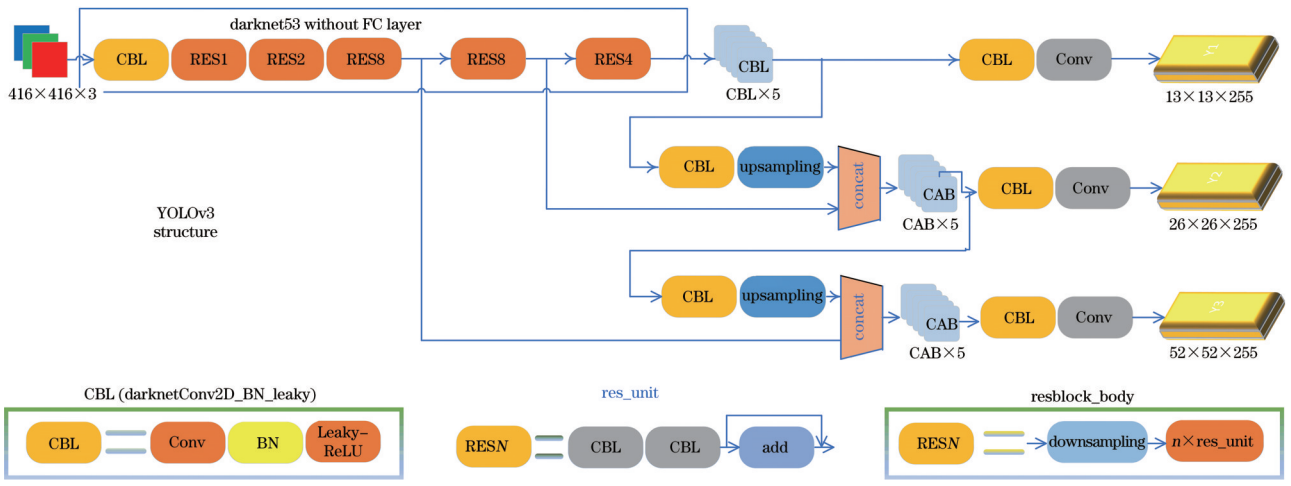

图 1　YOLOv3 网络结构图

Fig. 1　YOLOv3 network structure diagram

3　RDM-YOLOv3 模型

3.1　RD-Net 特征提取网络

DenseNet 通过构造一个身份映射来添加后续层的值，将所有层连接起来进行通道合并，以实现特征重用。与 ResNet 相比，DenseNet 提升了信息和梯度在网络中的传输效率，每层都能直接从损失函数得到梯度，并且直接得到输入信号，这样就能训练更深的网络，这种网络结构还有正则化的效果。其他网络致力于从深度和宽度来提升网络性能，DenseNet 致力于从特征重用的角度来提升网络性能。DenseNet 结构图如图 2 所示。

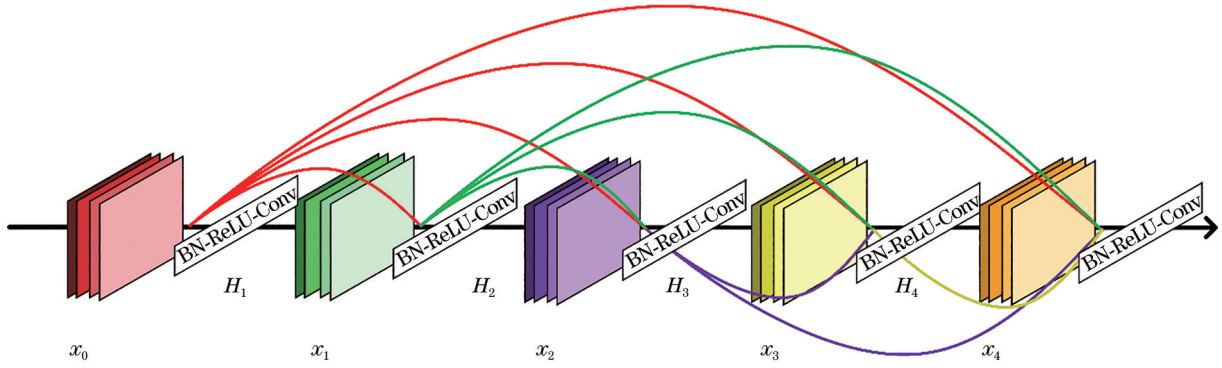

图 2　DenseNet 网络结构图

Fig. 2　DenseNet network structure diagram

DenseNet 和 ResNet 的一个明显区别是 ResNet 是求和，而 DenseNet 是进行拼接，每一层网络的输入包括前面所有层网络的输出。在图 1 中，x_1, x_2, x_3, x_4 代表输出的特征图，H_1, H_2, H_3, H_4 则代表非线性变换。普通的卷积 L 层有 L 个连接，DenseNet 的 L 层有 $L(L+1)/2$ 个连接，每一层都连接到所有其他层。因此，每一层可以接收先前的 $L-1$ 层的所有特征图。每层的特征图可以表示为

$$x_l = H_l[x_0, x_1, x_2, \ldots, x_{l-1}]。\tag{3}$$

本节提出的 DenseBlock 中的 CBL 模块是由卷积、批归一化和 Leaky-ReLU 构成。两个 CBL 模块构成了 D-CBL 模块，D-CBL 中使用 1×1 卷积来压缩通道数，使用 3×3 卷积扩张通道数，即 D-CBL 模块作为传输层。具体的传输层详细信息如表 1 所示。

DenseNet 层数过多会导致特征图变得多余并

表 1　传输层的内部通道信息

Table 1　Internal channel information of transport layer

Network	D-CBL of DenseBlock1	D-CBL of DenseBlock2
Structure	Conv (1×1×32)	Conv (1×1×64)
	BN	BN
	Leaky-ReLU	Leaky-ReLU
	Conv (3×3×64)	Conv (3×3×128)
	BN	BN
	Leaky-ReLU	Leaky-ReLU

降低检测速度，本节为每个模块设置了三层。DenseBlock1和DenseBlock2的结构如图3所示，RD模块主要将Residual Block和DenseBlock模块进行了整合，形成了Residual Dense Block。连续内存将F_{d-1}，$F_{d,1}F_{d,c}$，…，F_d等多层特征在同一纬度串联起来得到

$$L_{d,c}=\sigma\left(W_{d,c}\left[L_{d-1},L_{d,1},\cdots,L_{d,c-1}\right]\right),\tag{4}$$

式中：$W_{d,c}$代表第d个RD结构中第c个卷积操作；L_{d-1}代表第$d-1$个RD的输出；$\left[L_{d-1},L_{d,1},\cdots,L_{d,c-1}\right]$代表前$c-1$个卷积的输出。连续内存的密集连接保证了连续的低级高级信息的存储和记忆，每一个RD模块的输出都与上一个RD模块的输出级联。DenseBlock1的特征图增量为64，DenseBlock2的每一层特征图增量为128，自身映射连接所有通道数之后，使用1×1卷积的压缩层来压缩通道数，以便连接后续的网络。图3中$L_{d,al}$代表经过了3层RD模块的输出与L_{d-1}进行级联的结果。

RD-Net的具体结构如图4所示，在第三次残差模块后接入DenseBlock1，特征图由$T_1\in\mathbb{R}^{W\times H\times 512}$

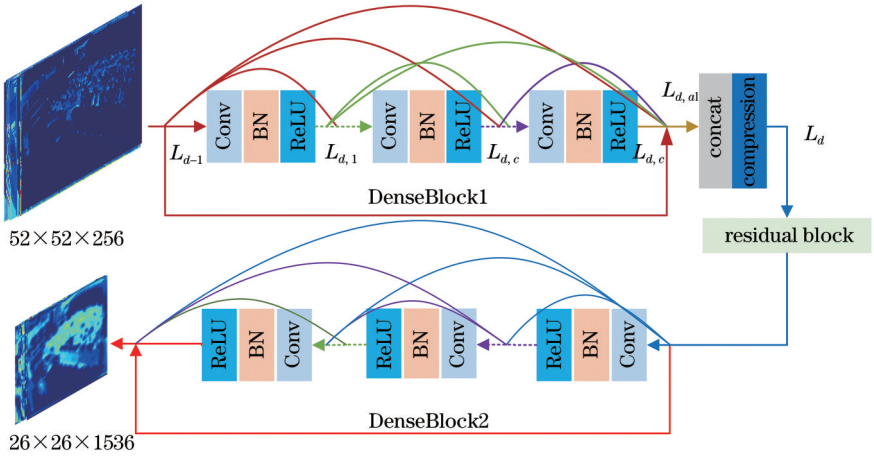

图3　DenseBlock内部结构图

Fig. 3　DenseBlock internal structure diagram

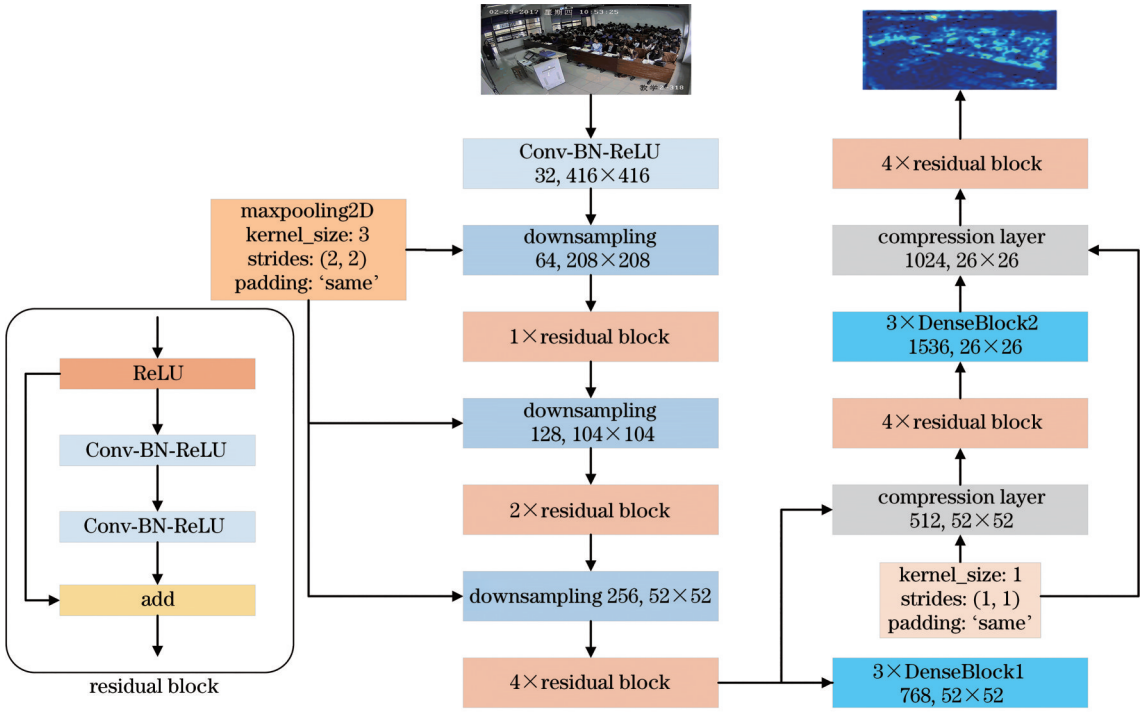

图4　RD-Net特征提取网络的结构

Fig. 4　Structure of RD-Net feature extraction network

变为 $T_2\in\mathbb{R}^{W\times H\times 768}$(其中 W 和 H 分别表示特征图的宽度和高度),然后经过压缩层再与 residual block 进行连接。在第 4 个残差模块后接入 DenseBlock2,特征图由 $T_3\in\mathbb{R}^{W\times H\times 1024}$ 变为 $T_4\in\mathbb{R}^{W\times H\times 1536}$,再与最后一个残差块连接。图 4 中列出了原始图像的特征提取具体参数变化(默认输入为 416 × 416)。

3.2　混合空洞卷积

虽然人的头部和肩部区域形状变化最小、稳定性高,但对于单幅图像来说,头部相对于其他物体来说是一个小目标,像素占比较小。在 YOLOv3 的 FPN 部分完成特征融合之后,进一步使用混合空洞卷积来融合上下文语义信息,增大特征层的感受野,以便更好地检测小目标。混合空洞卷积因膨胀率堆叠会有栅格效应,导致有的像素缺失,损失信息的连续性。膨胀率叠加需要满足

$$M_i=\max\left[M_{i+1}-2r_i,M_{i+1}-r_i,r_i\right],\qquad(5)$$

式中:r_i 是第 i 层的膨胀率;M_i 是第 i 层的最大膨胀率。假设共有 n 层,那么要满足 $M_n=r_n$[21]。利用 YOLOv3 的三个检测层,本文分别设置三个混合空洞卷积结构置于空间特征金字塔之后以提取不同感受野的特征图信息,然后融合输入与空洞卷积的信息,提取细粒度的语义信息。YOLOv3 有三个不同尺度的检测层,分别对应大、中、小物体的检测。在满足(5)式的条件下,本文在 13 × 13 检测层嵌入 MDC1,在 26 × 26 和 52 × 52 检测层嵌入 MDC2,完成混合空洞卷积的构建。MDC1 和 MDC2 具体结构如图 5 所示,三个不同尺度的 MDC1 的空洞卷积膨胀率 r 分别为 1, 2, 4, MDC2 的空洞卷积膨胀率 r 分别为 3, 4, 5。

本节以第一次上采样并完成空间特征金字塔融合之后的第一个 MDC2 为例,如图 6 所示,该模块中

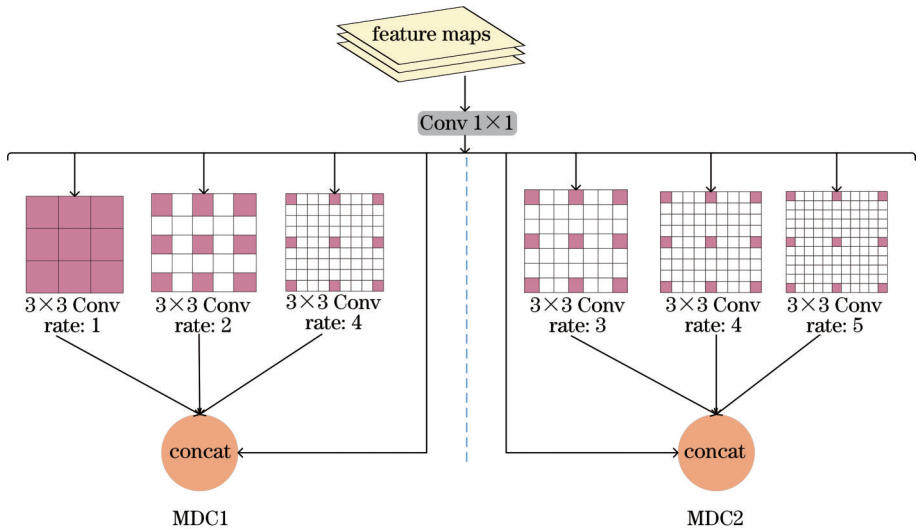

图 5　MDC 结构图

Fig. 5　MDC structure diagram

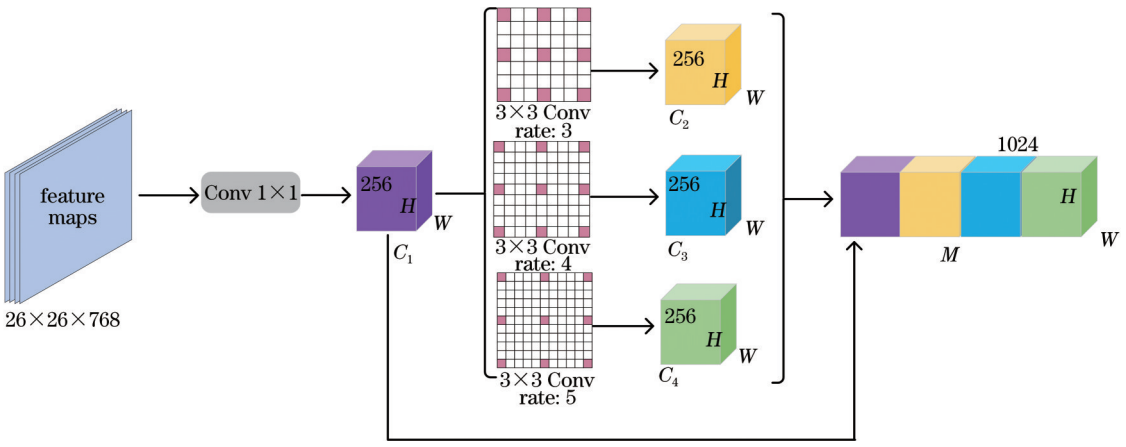

图 6　第一个 MDC2 通道连接图

Fig. 6　First MDC2 channel connection diagram

并行使用了3个不同膨胀率的空洞卷积，且膨胀率大小反映了相应的接收场大小。首先，通过一个1×1卷积把特征融合后的特征图$C\in\mathbb{R}^{W\times H\times 768}$的通道数压缩并获得特征图$C_1\in\mathbb{R}^{W\times H\times 256}$。然后使用不同膨胀率($r$=3,4,5)的空洞卷积在特征图$C_1$上采样，获得$C_2\in\mathbb{R}^{W\times H\times 256}$，$C_3\in\mathbb{R}^{W\times H\times 256}$，$C_4\in\mathbb{R}^{W\times H\times 256}$。最后将$C_1$，$C_2$，$C_3$，$C_4$进行融合，得到特征图$M\in\mathbb{R}^{H\times W\times 1024}$，$M=\left[C_1,C_2,C_3,C_4\right]$。通过该模型可以接收到不同感受野的信息，然后完成融合，以便提取更多的语义信息，尤其是针对小目标的特征信息。

本文将3个MDC模块积嵌入到YOLO-head前，完成空间特征金字塔的连接之后，使用MDC对4个上下文信息感知模块进行特征融合，再将融合特征送入YOLO-head进行检测。所提出的RDM-YOLOv3如图7所示。通过RD-Net提取图片特征，得到语义信息较为丰富的特征图，但是目标位置定位不准确，因此通过上采样把深层语义信息与浅层语义信息进行融合，可获得细节信息和丰富的语义信息。然后由MDC模块对特征图用不同膨胀率的空洞卷积层进行采样并融合，以扩大感受野，从而利用了更多的细粒度特征信息。

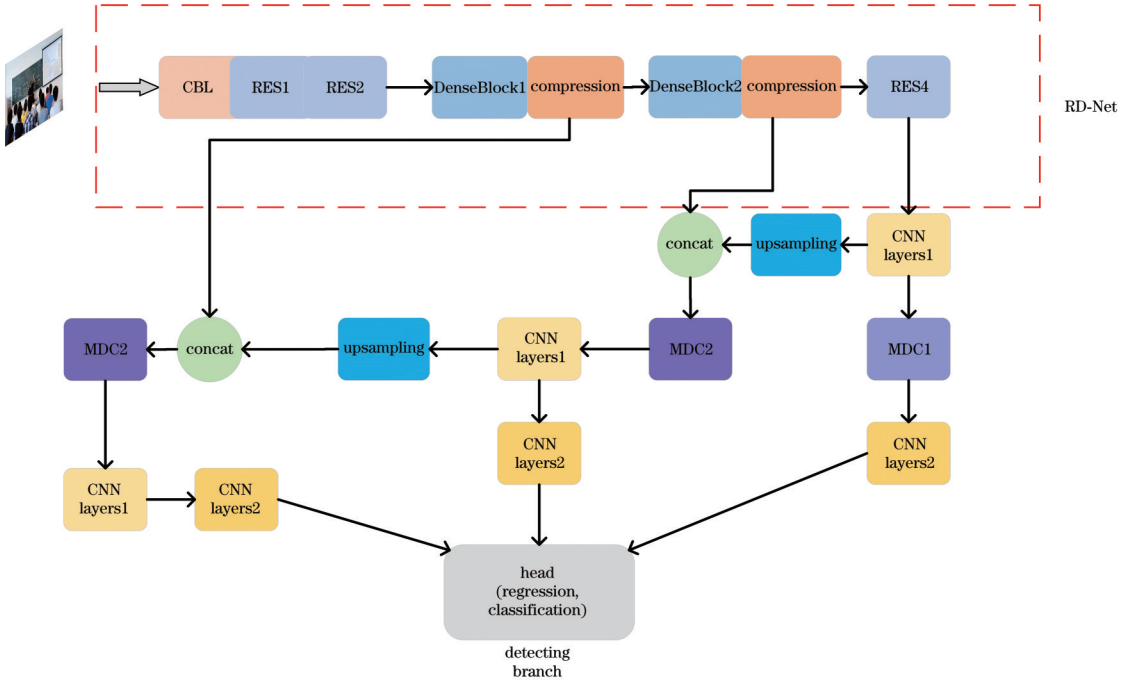

图7　完整的网络结构图

Fig. 7　Complete network structure diagram

4　实验条件与评估方法

本文所有的实验都在NVIDIA Tesla P100 GPU服务器上搭建的Pytorch环境下进行，显存为16G，操作系统为Ubuntu 7.5.0。

为公平起见，本文实验结果的评估使用AP(average precision)，本文采取IOU阈值为0.5作为绘制precision-recall(PR)曲线的基准。PR曲线与坐标围成的面积表示AP值，AP值越大表示模型精度越高，性能越好，AP的计算公式为

$$V_{AP}=\frac{1}{N}\sum_{i}^{N}\frac{N_{TP}}{N_{TP}+N_{FP}},\tag{6}$$

式中：N_{TP}为真阳性，即预测为正、实际也为正的样本数(真正例)；N_{FP}为假阳性，即预测为正、实际为负的样本数(假正例)。本文选取0.5作为特定的IOU阈值，当预测边框的IOU大于等于0.5，则认为可以正确检测到头部[22]。在该标准下可以获得Precision、Recall的值，其中Precision(P)表示在识别出来的图片中，正样本所占的比率，Recall(R)表示正样本被预测为正类图片占所有测试集图片的比率。总体而言，本文采用三个标准指标[P、R和AP-IOU(用$M_{AP\text{-}IOU}$表示)]来评估所提出的方法。各项评价指标的计算公式如下：

$$M_{AP\text{-}IOU}=\frac{N_{TP}}{N_{TP}+N_{FN}+N_{FP}},\tag{7}$$

$$P=\frac{N_{TP}}{N_{TP}+N_{FP}},\tag{8}$$

$$R=\frac{N_{TP}}{N_{TP}+N_{FN}},\tag{9}$$

式中：N_{FN}为假阴性，即预测为负、实际为正的样本数（假负例）。

5　实验结果与分析

5.1　消融实验

如上文所述，本文提出了一个基于ResNet与DenseNet的特征提取网络RD-Net，以实现特征重用，使模型轻量化并提高层间信息透过率。此外，本文嵌入了MDC1和MDC2模块以与FPN融合，该结构位于YOLO head前，其目的在于提取更多小目标的信息，提升算法对头部检测的准确率。为了证明本文提出方法的有效性，进行了消融实验研究。首先用DarkNet-53（baseline）在两个数据集上进行实验，然后使用RD-Net进行对比实验，表2的实验结果表明本文提出的特征提取网络性能优于DarkNet-53，同时RD-Net减少了原来DarkNet-53的两个8×Residual Block的深层卷积参数，使用特征图自身的映射完成特征重用，这使得检测速度更快，平均FPS（Frames Per Second）提升了9。如图6所示，分别增加MDC1、MDC1+MDC2、MDC1+2MDC2进行消融实验。结果表明，与DarkNet-53相比，本文所提出方法的AP在HollywoodHeads、Brainwash测试集中分别提高了17.9%和16.3%。图8中的PR曲线和表2详细展示了本文方法与基线DarkNet-53的测试对比，通过消融实验可知：RD-Net+MDC1+2MDC2的性能更佳，其PR曲线面积更大，平滑性更好；RDM-YOLOv3模型的检测稳定性更好，这证明了该方法的有效性。

表2　两个数据集上的消融实验比较

Table 2　Comparison of ablation experiments on two data sets

Method	AP (R_{IOU}=0.5)		FPS
	HollywoodHeads	Brainwash	
DarkNet-53 (baseline)	0.689	0.752	19
RD-Net	0.750	0.801	28
RD-Net+MDC1	0.782	0.846	25
RD-Net+MDC1+MDC2	0.823	0.889	21
RD-Net+MDC1+2×MDC2	0.868	0.931	16

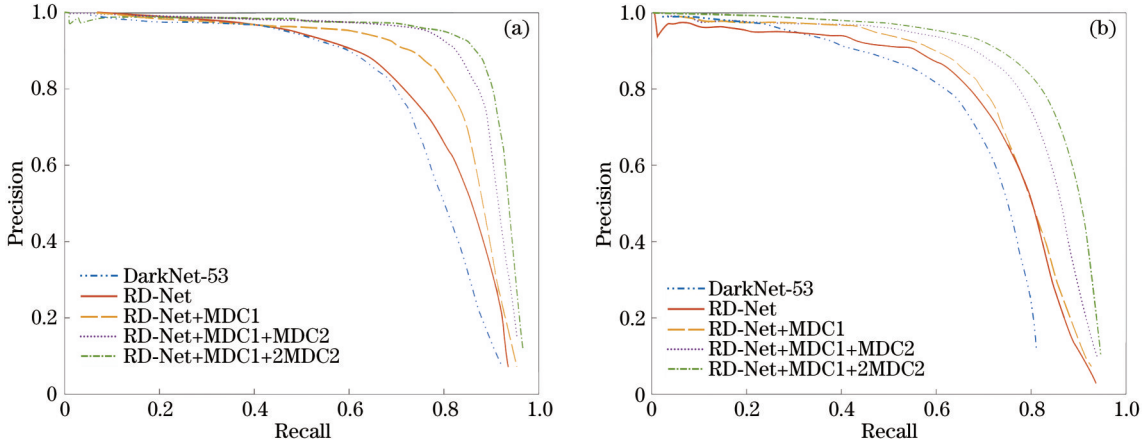

图8　两个数据集上的消融实验的PR曲线。(a) Brainwash数据集；(b) HollywoodHeads数据集

Fig. 8　PR curves of ablation experiments on two datasets. (a) Brainwash dataset; (b) HollywoodHeads dataset

5.2　两个头部数据集实验结果

Brainwash数据集的所有图像都是从一间咖啡店监控摄像头中剪裁的，并且人员分布非常密集。将本文中提出的方法与几个代表性的检测器进行了比较，如FCHD[23]、E2PD[8]、SSD[13]、HeadNet[24]、FRCN[25]等。可以看出，本文提出的方法的性能比其他方法更好，在R_{IOU}为0.5时，本文方法的AP比YOLOv3提升了约21.2%；E2PD使用的长短期记忆网络LSTM作为Backbone，本文的RD-Net+MDC更加侧重在小目标上，在该人头数据集中性能优于E2PD，相比目前效果最好的HeadNet，在Brainwash测试集上的AP值提升了约2.3%。在Brainwash测试集上的具体结果对比图如表3和图9所示。

在HoolywoodHeads数据集上，本文使用了6种不同方法进行对比实验。其中包括基于CNN的通用目标检测方法FRCN、SSD；考虑到头部检测类似于人脸检测，本文选取一个基于CNN的面部检测器DPM-Face[26]。本文还将RDM-YOLOv3与使用

表 3　Brainwash 数据集上的比较结果

Table 3　Comparison results on Brainwash dataset

Method	Backbone	AP (R_{IOU}=0.5)
SSD	VGG16	0.568
FCHD	VGG16	0.700
YOLOv3	DarkNet-53	0.768
E2PD	GoogLeNet+LSTM	0.821
FRCN	VGG16	0.878
HeadNet	ResNet-101	0.910
RDM-YOLOv3	RD-Net+MDC	0.931

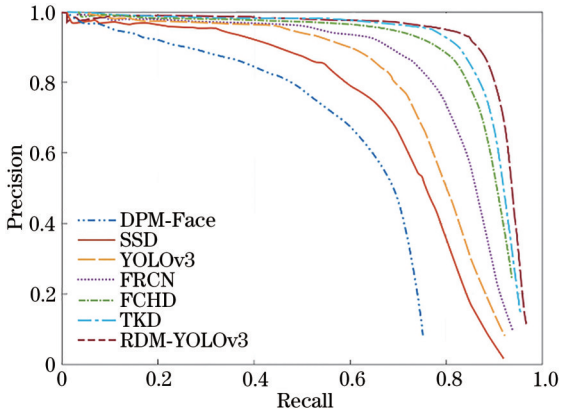

图 9　RDM-YOLOv3 与其他方法在 Brainwash 数据集上的 PR 曲线比较

Fig. 9　Comparison of PR curves between RDM-YOLOv3 and other methods on Brainwash dataset

LSTM 的 TKD[27]的新方法进行了比较，实验结果表明本文所提出的模型比 TKD 性能更好。表 4 中 DPM-Face 检测率最低，其原因是使用了多任务级联卷积网络进行人脸检测和对齐，该数据集中存在大量侧脸和头部背面，导致检测率过低。本文方法在 HollywoodHeads 数据集上的 AP 值为 0.868，优于其他方法。表 4 和图 10 展示了详细的对比信息。

表 4　HollywoodHeads 数据集上的比较结果

Table 4　Comparison results on HollywoodHeads dataset

Method	Backbone	AP (R_{IOU}=0.5)
DPM-Face	—	0.370
SSD	VGG16	0.621
YOLOv3	DarkNet-53	0.689
FRCN	VGG16	0.712
FCHD	VGG16	0.743
TKD	LSTM	0.750
RDM-YOLOv3	RD-Net+MDC	0.868

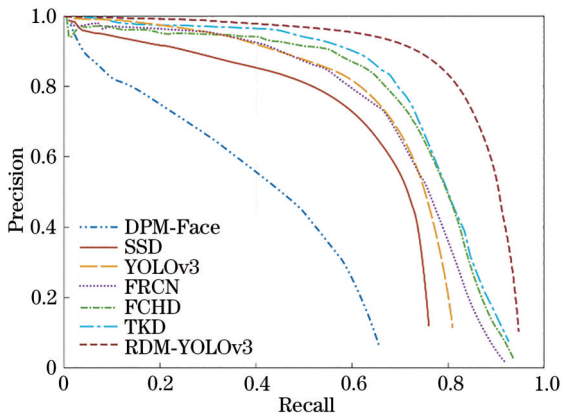

图 10　RDM-YOLOv3 与其他方法在 HollywoodHeads 数据集上的 PR 曲线比较

Fig. 10　Comparison of PR curves between RDM-YOLOv3 and other methods on HollywoodHeads dataset

5.3　测试结果对比

本文所提方法和 YOLOv3 在两个公开数据集上的一些可视化结果如图 11 所示，图中第一列是 YOLOv3 的检测效果图，第二列是 RDM-YOLOv3 的检测效果图。在 Brainwash 数据集中，人员较为密集，YOLOv3 存在误检和漏检，RDM-YOLOv3 在人员密集场景中的检测效果优于 YOLOv3；在 HollywoodHeads 数据集中，测试图片存在动态模糊

图 11　两个头部数据集测试对比图。(a) HollywoodHeads 数据集；(b) Brainwash 数据集

Fig. 11　Comparison of test results of two head datasets. (a) HollywoodHeads dataset; (b) Brainwash dataset

和失真的情况，YOLOv3对于模糊的人头有漏检的情况，而RDM-YOLOv3能够完成检测。从这两幅图中可以看出RDM-YOLOv3的漏检率和误检率比YOLOv3低。测试结果展示了本文提出的RDM-YOLOv3对小目标的检测率更高。

6 结 论

在YOLOv3的基础上提出了小目标检测性能更好的模型RDM-YOLOv3。为了减少参数量和提高层之间的信息透过率，提出了一种基于DenseNet和ResNet的特征提取网络RD-Net，从而获得更多语义信息。目前的大多数神经网络在小目标检测上仍然存在漏检问题，提出了通过在FPN之后嵌入不同采样率的空洞卷积构成的MDC1和MDC2结构，以提高对头部检测的敏感度。通过消融实验证明了该方法的有效性。在两个具有挑战性的数据集上进行了广泛的实验，实验结果表明，所提方法显著提高了头部检测的准确性。该方法的局限是当场景中头部密集度很高时，如体育馆、景区等场景，仍存在漏检的问题。下一步计划是使用Transformer替代CNN，引入多头注意力机制，以更加有效地提升小目标的检测率。此外，人群计数是头部检测的一个重要应用，未来将在密集人群进行检测和计数，将头部检测应用于人群密集场景中，以实现超过人流阈值报警等应用。

参 考 文 献

[1] Li N, Wu Y Y, Liu Y, et al. Pedestrian attribute recognition algorithm based on multi-scale attention network[J]. Laser & Optoelectronics Progress, 2021, 58(4): 0410025.
李娜, 武阳阳, 刘颖, 等. 基于多尺度注意力网络的行人属性识别算法[J]. 激光与光电子学进展, 2021, 58(4): 0410025.

[2] Tian Y C, Dehghan A, Shah M. On detection, data association and segmentation for multi-target tracking [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(9): 2146-2160.

[3] Yao H T, Zhang S L, Hong R C, et al. Deep representation learning with part loss for person re-identification[J]. IEEE Transactions on Image Processing, 2019, 28(6): 2860-2871.

[4] Yu C Y, Xu Y, Gou L S, et al. Crowd counting based on single-column deep spatiotemporal convolutional neural network[J]. Laser & Optoelectronics Progress, 2021, 58(8): 0810011.
鱼春燕, 徐岩, 缑丽莎, 等. 基于单列深度时空卷积神经网络的人群计数[J]. 激光与光电子学进展, 2021, 58(8): 0810011.

[5] Zhang T, Zhang L. Multiscale feature fusion-based object detection algorithm[J]. Laser & Optoelectronics Progress, 2021, 58(2): 0215003.
张涛, 张乐. 一种基于多尺度特征融合的目标检测算法[J]. 激光与光电子学进展, 2021, 58(2): 0215003.

[6] Ballotta D, Borghi G, Vezzani R, et al. Fully convolutional network for head detection with depth images[C]//2018 24th International Conference on Pattern Recognition (ICPR), August 20-24, 2018, Beijing, China. New York: IEEE Press, 2018: 752-757.

[7] Shami M B, Maqbool S, Sajid H, et al. People counting in dense crowd images using sparse head detections[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2019, 29(9): 2627-2636.

[8] Zhang J J, Liu Y T, Li R C, et al. End-to-end spatial attention network with feature mimicking for head detection[C]//2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), November 16-20, 2020, Buenos Aires, Argentina. New York: IEEE Press, 2020: 199-206.

[9] Vu T H, Osokin A, Laptev I. Context-aware CNNs for person head detection[C]//2015 IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015, Santiago, Chile. New York: IEEE Press, 2015: 2893-2901.

[10] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2014, Columbus, OH, USA. New York: IEEE Press, 2014: 580-587.

[11] Gao S H, Cheng M M, Zhao K, et al. Res2Net: a new multi-scale backbone architecture[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(2): 652-662.

[12] Hariharan B, Arbeláez P, Girshick R, et al. Hypercolumns for object segmentation and fine-grained localization[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE Press, 2015: 447-456.

[13] Liu W, Anguelov D, Erhan D, et al. SSD: single shot MultiBox detector[M]//Leibe B, Matas J, Sebe N,

et al. Computer vision-ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016, 9905: 21-37.

[14] Redmon J, Divvala S, Girshick R, et al. You only look once: unified, real-time object detection[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 779-788.

[15] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 770-778.

[16] 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), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 936-944.

[17] Cheng B W, Wei Y C, Shi H H, et al. Revisiting RCNN: on awakening the classification power of faster RCNN[M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11219: 473-490.

[18] Raj R J S, Shobana S J, Pustokhina I V, et al. Optimal feature selection-based medical image classification using deep learning model in internet of medical things[J]. IEEE Access, 2020, 8: 58006-58017.

[19] Huang G, Liu Z, van der Maaten L, et al. Densely connected convolutional networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 2261-2269.

[20] Fan Q, Zhuo W, Tang C K, et al. Few-shot object detection with attention-RPN and multi-relation detector[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 4012-4021.

[21] Wang P Q, Chen P F, Yuan Y, et al. Understanding convolution for semantic segmentation [C]//2018 IEEE Winter Conference on Applications of Computer Vision (WACV), March 12-15, 2018, Lake Tahoe, NV, USA. New York: IEEE Press, 2018: 1451-1460.

[22] Everingham M, Eslami S M A, Gool L, et al. The pascal visual object classes challenge: a retrospective [J]. International Journal of Computer Vision, 2015, 111(1): 98-136.

[23] Vora A, Chilaka V. FCHD: fast and accurate head detection in crowded scenes[EB/OL]. (2018-09-24) [2021-05-20]. https://arxiv.org/abs/1809.08766v3.

[24] Li W, Li H L, Wu Q B, et al. HeadNet: an end-to-end adaptive relational network for head detection[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 30(2): 482-494.

[25] 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.

[26] Ranjan R, Patel V M, Chellappa R. A deep pyramid deformable part model for face detection[C]//2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS), September 8-11, 2015, Arlington, VA, USA. New York: IEEE Press, 2015: 1-8.

[27] Farhadi M, Yang Y Z. TKD: temporal knowledge distillation for active perception[C]//2020 IEEE Winter Conference on Applications of Computer Vision (WACV), March 1-5, 2020, Snowmass, CO, USA. New York: IEEE Press, 2020: 942-951.