融合轻量化与梯形结构的学生行为检测算法*

张 颖[1],张 喆[1],龙光利[2]
(1.西安理工大学 自动化与信息工程学院,陕西 西安 710048;
2.陕西理工大学 物理与电信工程学院,陕西 汉中 723000)

摘 要:为了解决常见目标检测算法在课堂场景中难以有效应用的问题,提出了一种融合轻量化与梯形结构的学生行为检测算法。该算法基于YOLOv4架构,针对目标分类和分布空间的特点,提出一种新的“梯”形特征融合结构,并结合MobileNetv2思想,优化模型参数得到梯形-MobileDarknet19特征提取网络,既减少了网络的计算量,提高了工作效率,同时加强了目标特征的信息传输,提升了模型学习能力;在尺度检测阶段引入5层的DenseNet网络,增强网络对小目标的检测能力。实验结果表明,提出的YOLOv4-ST算法相比于原YOLOv4算法mAP提高了5.5%,相比于其他主流算法,在学生课堂行为检测任务中具有较好的实用性。

关键词:梯形结构;学生行为检测;YOLOv4;特征融合;DenseNet
中图分类号:TP399 **文献标识码**:A DOI:10.16157/j.issn.0258-7998.222837

中文引用格式:张颖,张喆,龙光利. 融合轻量化与梯形结构的学生行为检测算法[J].电子技术应用,2022,48(12):47-53.
英文引用格式:Zhang Ying,Zhang Zhe,Long Guangli. Student behavior detection algorithm combining lightweight and trapezoidal structure[J]. Application of Electronic Technique,2022,48(12):47-53.

Student behavior detection algorithm combining lightweight and trapezoidal structure

Zhang Ying[1],Zhang Zhe[1],Long Guangli[2]
(1.School of Automation and Information Engineering,Xi′an University of Technology,Xi′an 710048,China;
2.School of Physics and Telecommunications Engineering,Shaanxi University of Technology,Hanzhong 723000,China)

Abstract: In order to solve the problem that common target detection algorithms are difficult to apply effectively in classroom scenarios, a student behavior detection algorithm combining lightweight and trapezoidal structure is proposed. The algorithm is based on YOLOv4 architecture, according to the characteristics of target classification and distribution space, a new "trapezoidal" feature fusion structure is proposed, and combined with the MobileNetv2 idea, the model parameters are optimized to obtain a trapezoidal-MobileDarknet19 feature extraction network, which not only reduces the computational load of the network, but also improves the work efficiency. At the same time, it strengthens the information transmission of target features and improves the learning ability of the model. In the scale detection stage, a five-layer DenseNet network is introduced to enhance the network′s detection ability for small targets. The experimental results show that the proposed YOLOv4-ST algorithm is better than the original one. The mAP of YOLOv4 algorithm is improved by 5.5%. Compared with other mainstream algorithms, it has better practicability in the task of student classroom behavior detection.

Key words: trapezoidal structure; student behavior detection; YOLOv4; feature fusion; DenseNet

0 引言

随着教育现代化的推进,信息化教学越来越普遍,作为学校教育中最基本也是最重要的环节,课堂教学面临着传统走向现代的变革。课堂中,老师通过观察学生的表现获得授课情况的反馈,但一对多的教学方式存在着观察不全面、可信度低、无法实时掌握学生课堂学习情况等问题。将基于深度学习的行为检测应用到课堂教学场景中,通过实时监控,帮助老师全面地掌握学生课堂状态,及时合理地调整教学进度和策略,不仅能够提高教学效率,还能够推动智能化教学的发展,为今后现代化课堂的探索奠定了基础。

近年来,随着深度学习在计算机视觉领域取得了突破性进展[1],Faster-RCNN(Faster Region-based Convolutional Neural Networks)[2]、SSD(Single Shot Detection)[3]、YOLO(You Only Look Once)[4-6]等目标检测算法也相继出现。Zheng等[7]通过一种新的特征融合策略改进Faster R-CNN进行行为检测,但检测精度不高;Liu等[8]提出了一种基于双流结构的改进时空注意力模型,将空间和时间特征

*基金项目:国家自然科学基金(61971345);陕西省重点产业创新链工程(2020ZDLGY05-02)

分别馈入空间长短期记忆(Long Short-rm Memory,LSTM)和时间 LSTM,融合特征来识别视频中的不同动作;2020 年,Bochkovskiy 等[9]提出 YOLOv4 算法,该算法的网络骨干结构使用了结合跨阶段部分连接[10](Cross Stage Partial Connection)与 Darknet53 结合而形成的 CSPDarknet53 特征提取结构,有效提升了检测精度和速度;Ren 等[11]通过在 YOLOv4 的特征提取结构中添加了跳跃式的连接,能够融合更多的特征,在一定程度上提升了学生行为检测精度,但效率较低。以上研究表明,深度学习用于学生行为检测具有一定的理论基础和实践可行性。虽然许多检测算法在应用领域表现优异,但对于课堂场景来说,学生活动范围有限,受摄像头位置及视觉角度的影响,学生目标较小且行为易受遮挡,导致会出现漏检错检、检测精度低等问题。其次,课堂学生行为检测需要建立特定的学生行为数据集,要从海量的课堂监控视频进行筛选和制作,并选用适合的先验框参数,以适应学生目标尺寸。

根据上述问题及难点,本文提出了一种改进的 YOLOv4-ST 算法实现对学生课堂异常行为的检测。首先,采集课堂监控视频构建学生行为数据集,观察选取了睡觉、玩手机、交头接耳 3 种常见课堂异常行为进行标注,并提出了一种新的“梯”形特征提取结构:梯形-MobileDarknet19,将不同尺度的特征进行融合,提升网络抓取图像的局部特征和低频信息的能力,增强图像的上下文语义信息;此外,在小尺度检测阶段加入了密集连接结构,加强了特征的传输能力,提高了模型对小目标的检测精度。实验结果表明,改进后的 YOLOv4 算法对学生课堂异常行为检测的速度和精度均优于其他算法,可以更好地满足实际应用需求。

1 YOLOv4 算法原理

YOLOv4 算法的主要思路是将目标检测问题转化成一个回归问题,当输入图像后,采用单个神经网络来回归输出目标边界与其类别概率,以实现端到端的直接预测。网络结构如图 1 所示。

CSPDarknet53 作为 YOLOv4 算法提取目标特征的核心,借鉴了 CSPNet 网络[10]的原理,将特征映射划分为两个部分,通过跨阶段合并,有效增加了网络的学习能力,降低计算量。激活函数选取了 Mish 函数[12],避免模型出现饱和的问题,使模型有更好的泛化能力。CSPDarknet53 网络之后引入了空间金字塔池化(Space Pyramid Pool,SPP)[13]模块和路径聚合网络(Path Aggregation Network,PAN)[14]的结构,将尺寸为 19×19 的特征图送入 SPP 网络中,进行不同尺度的池化操作,提高了感受野,增加了主干特征的接收范围,有效分离了上下文特征[15]。得到的特征图经过堆叠、卷积后输出到 PANet 中,PANet 结构是在特征金字塔(Feature Pyramid Network,FPN)[16]基础上又添加了一个自底向上的倒金字塔结构,避免了浅层信息的丢失,增强了特征的传输能力。最后预测层根据先验

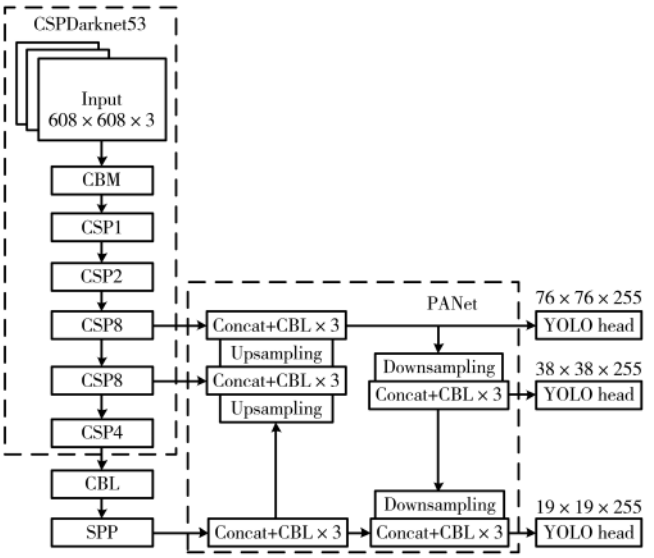

图 1 YOLOv4 网络结构

框参数、目标置信度、类别概率等信息,以极大值抑制算法选取最佳的预测框[17]。

2 改进的 YOLOv4-ST 算法

2.1 特征提取网络轻量化

神经网络通常通过加深模型层数来获得优异的性能,CSPDarknet53 结构的使用虽然提升了网络性能,但令网络结构更加复杂化。根据本文所采集的课堂环境下样本数量和检测类别的特点,需要对特征提取网络结构进行优化。考虑到本文所检测类别仅限定为 3 种,借鉴 MoblieNetv2[18]思想,提出 MoblieDarknet19,通过引入深度可分离卷积代替传统卷积方式,从而降低原特征提取结构的参数冗余度,在保持同样检测精度的前提下提升了模型工作效率。MoblieDarknet19 参数如表 1 所示。

表 1 MoblieDarknet19 结构参数

类型	步进	数量 r	输出
Conv	–	1	304×304×32
Bot1	1	1	304×304×16
Bot2	2	2	152×152×24
Bot2	2	3	76×76×32
Bot2	2	4	38×38×64
Bot1	1	3	38×38×96
Bot2	2	3	19×19×160
Bot1	1	1	19×19×320
Conv	–	1	19×19×1280

由表 1 可知,MoblieNet19 包含了 Bot1 和 Bot2 结构,两者主要区别在于中间使用的 DepthWise 层的卷积步长不同。文中采用 DepthWsie 和 PointWsie 层分别代替传统卷积层来提取图像特征,可以有效地减少模型的时间及空间复杂度,图 2 是改进后的结构示意图。其中,第一块 PointWise 结构由 1×1 卷积、BN 层和 ReLU6 组成,该模块

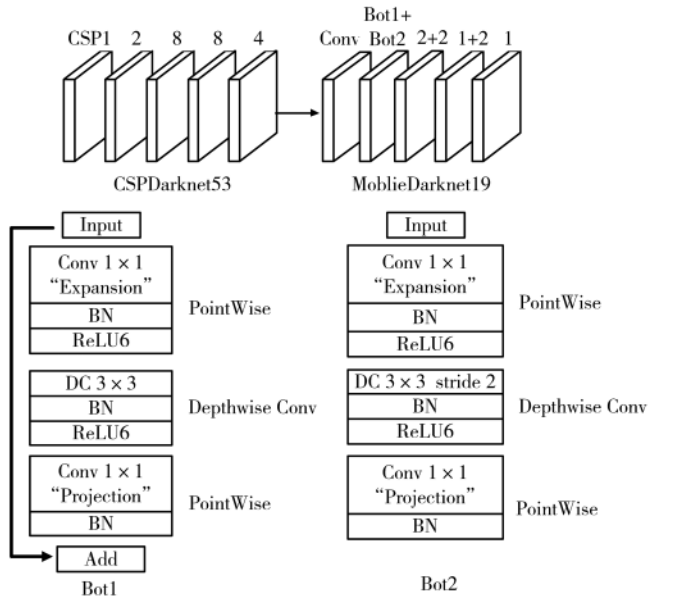

图 2 MobleDarknet19 结构图

将目标特征从低维空间映射到高维空间，有利于模型对特征的提取和学习；第二块 DepthWise 结构由 3×3 卷积、BN 层和 ReLU6 组成；第三块 PointWise 结构则将特征从高维空间映射到了低维空间，并且由于 ReLU6 会破坏模型从低维空间学习到的特征参数，因此最后舍弃了该激活函数。

2.2 "梯"型特征融合结构

同时，由于课堂场景中学生密集且存在遮挡的情况，提出一种新的特征融合方式——"梯"形结构。在 MobleDarknet19 特征提取框架网络中的 152×152、76×76 和 38×38 特征图中分别引入卷积层，形成"梯"形结构，并将所得特征图与输出的相同尺度大小特征图进行拼接，以融合不同层次的特征，如图 3 所示。

梯形特征融合具有以下优点：(1)深层特征通常具有高语义信息，浅层特征则包含了丰富的轮廓信息，因此加强与浅层特征层的结合，可以最大程度保留目标细节信息，进一步提高对小目标的检测能力；(2)随着模型层数的增加，干扰的噪声信息含量也会增加，所以相对于深层信息，浅层信息噪声携带量少，特征信息更纯粹，梯形特征融合结构加强深浅结合，有效增强了图像特征传输能力；(3)梯形特征融合方式，增强了感受野较小的浅层特征的传输能力更加适合解决本文数据集存在的视距较远和遮挡情况导致目标较小的问题。

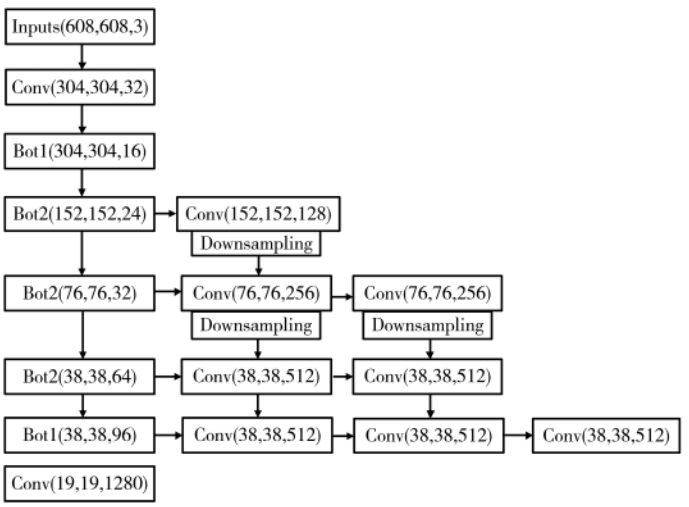

图 3 梯形-MobleDarknet19 结构图

2.3 融合 DenseNet 的预测阶段

YOLOv4 网络中，预测层根据不同大小、不同感受野的检测目标，输出 19×19、38×38、76×76 3 个不同尺度的预测框，由于网络中较多的池化和降采样操作，这 3 种尺度预测框对小目标检测精度不高。特别是，课堂场景下学生行为在整个图像中所占比例较小，使得检测难度增大。因此，本文在 76×76 的小检测尺度上引入了 5 层 DenseNet[19]网络，通过增加特征图的维度，充分利用浅层特征信息，加强小感受野的信息传递，提高模型对小目标的检测能力。具体结构如表 2 所示。图 4 为 YOLOv4-ST 网络模型图。

表 2 加入的 5 层 DenseNet 结构

类型	卷积核 r	大小	输出
Conv	128	1×1	76×76×128
Conv	256	3×3	76×76×256
Concat-1,-2	-	-	76×76×384
Conv	128	1×1	76×76×128
Concat-1,-2	-	-	76×76×512
Conv	256	3×3	76×76×256
Concat-1,-2	-	-	76×76×768
Conv	128	1×1	76×76×128
Concat-1,-2	-	-	76×76×896
Conv	256	3×3	76×76×256
Concat-1, -2	-	-	76×76×896
Conv	24	1×1	76×76×24

3 实验结果与分析

3.1 实验环境及数据集

本文是在 GPU Intel® Core™ i7-11700k CPU@3.60 GHz，Ubuntu18.04 系统下搭建的 Darknet 环境中进行的，显卡为 NVIDIA GeForce RTX 3060 12 GB 独立显卡，CUDA 和 cuDNN 版本为 10.1，Tensorflow 版本为 1.14，Python 版本为 3.6.4。

实验采集了西安理工大学教学楼的教室监控视频数据，所获数据来自于 5 个不同课堂不同视角的摄像头，视频总时长为 400 min。首先，筛选视频帧获得 1 855 张图片，数据集格式为 VOC2007，并按照 7∶2∶1 划分为训练集、验证集和测试集，数据集图片如图 5 所示。使用 LableImg 标注软件对训练集中玩手机、睡觉、交头接耳说话这 3 种出现频率较高的学生课堂异常行为进行标

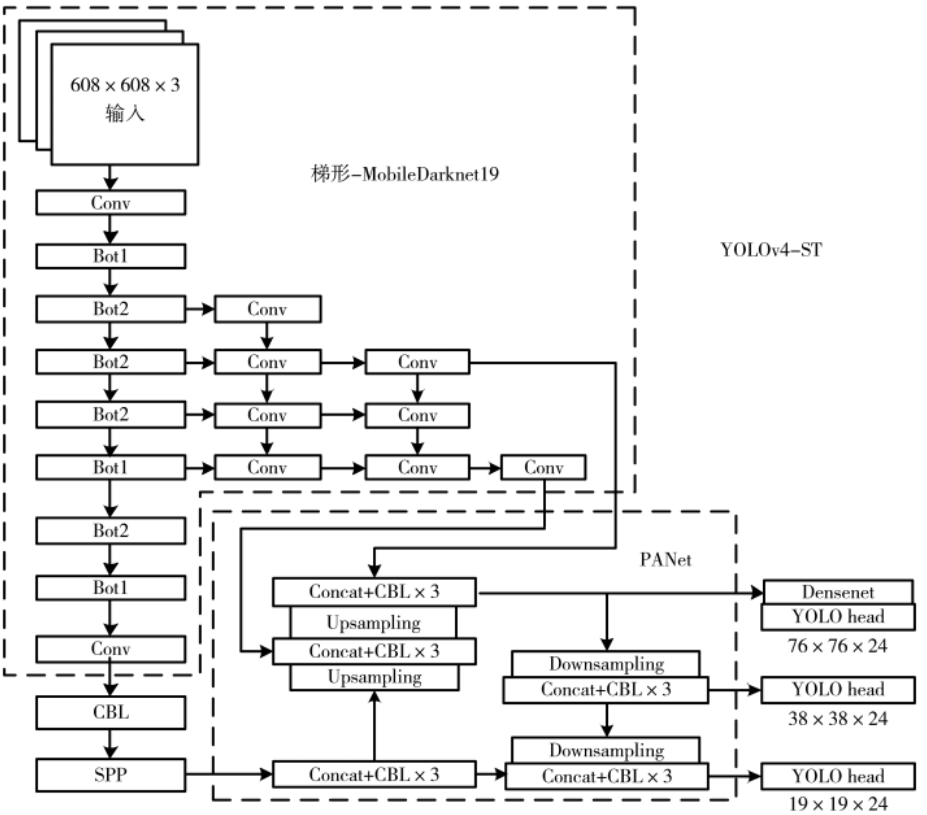

图 4　整体 YOLOv4-ST 网络模型图

图 5　部分数据集展示

注，标注类别分为 Playing phone、Sleeping 和 Talking。最后，将训练的模型参数在验证集和测试集中进行验证。

进行网络训练时，采用了小批量随机梯度下降法，批量大小为 64，最大迭代次数为 50 500 次，动量参数设置为 0.9，衰减系数设置为 0.000 5，初始化学习率为 0.01。当迭代 40 000 次时，调整学习率为 0.001；当迭代 45 000 次时，调整学习率为0.000 1。图 6 为 YOLOv4-ST 训练结果的 Loss 图，横坐标为迭代次数，纵坐标为训练 Loss 值。

3.2 消融实验及结果分析

为了验证本文算法的有效性，进行了 4 组消融实验。每组均在相同软硬件条件、相同数据增强方式以及训练策略下进行。消融实验如表 3 所示，实验 A 代表了原本的 YOLOv4 网络；实验 B 代表使用 MobileDarknet19 作为特征提取网络的网络模型；实验 C 代表在实验 B 的基础上引入“梯”形特征融合结构的网络模型；实验 D 基于实验 C，在预测阶段加入密集连接网络，即代表了本文提出的 YOLOv4-ST 网络模型。

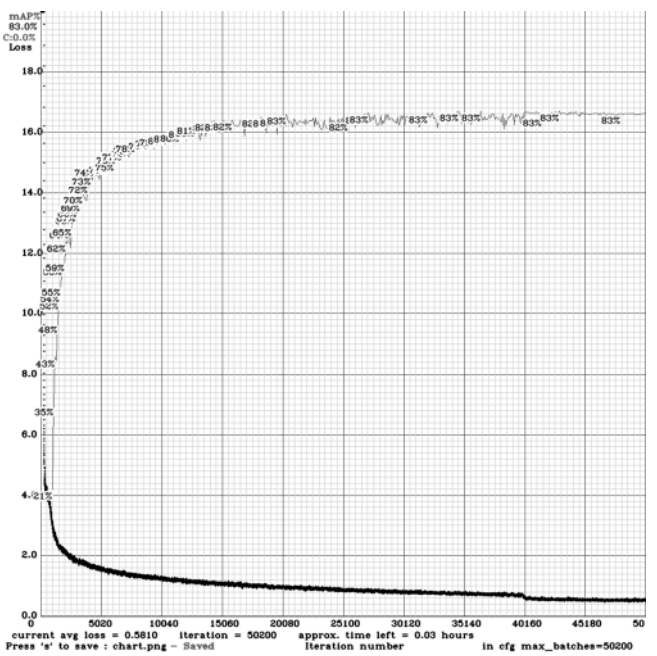

图 6　YOLOv4-ST 训练 Loss 图

从表 3 可以看出，使用 MobileDarknet19 作为特征提取

表 3 消融实验

方法策略	Mobile Darknet19	梯形特征融合方式	DenseNet	mAP/%	速率/fps	模型大小/MB
实验 A				77.5	28.7	244
实验 B	√			75.7	48.5	64.2
实验 C	√	√		81.3	43.7	68.7
实验 D	√	√	√	83.0	40.9	73.9

网络的实验 B 检测速率最高，达到48.5 fps，检测精度(mAP)降低了 1.8%。这是因为 YOLOv4 模型的计算量庞大，内存占用高达 244 MB，当训练样本和检测类别较少时，原 CSPDarknet53 网络中过多的网络参数造成了冗余，消耗了大量资源，因此训练和检测效率低下。本文设计的 MobileDarknet19 网络使得整个模型的内存占用降低到 64.2 MB，残差单元的有效减少，降低了网络的计算量，在网络内部，图像特征从低维映射到高维，有利于模型的学习过程，在保持检测精度的前提下提升了工作效率。实验 C 在 B 的基础上融入了“梯”形特征融合结构，虽然增加了 4.5 MB 的内存占用，但有效减少了细节信息的丢失，降低噪声传入，加强了目标特征传输，使得模型平均召回率提高了 2.3%，mAP 值达到了 81.3%。对比实验 B 和 C，融合“梯”形特征结构的网络模块虽然牺牲了少量的检测速率，但提升了检测精度，因此适用于本文所述应用场景的行为检测。实验 D 为本文所提出的 YOLOv4-ST 模型，其模型大小仅为原始 YOLOv4 模型的 1/3，在小尺度检测阶段加入了密集连接模块，通过短路连接的方式，提升了特征信息的传输效率，其 mAP 值达到 83.0%，提高了模型对学生异常行为的检测能力，同时满足实时检测的要求。

3.3 与其他模型的对比实验

为了验证 YOLOv4-ST 算法在检测精度和检测速率上的提升，本文在所建立的学生课堂行为数据集上与其他主流算法进行对比，结果如表 4 所示。

表 4 对比实验

模型	mAP/%	速率/fps
Faster-RCNN	80.6	20.4
SSD	79.2	24.3
YOLOv3	75.3	26.5
YOLOv4	77.5	28.7
YOLOv4-tiny	71.3	60.9
YOLOv4-ST	83.0	40.9
参考文献[20]	79.5	36.5

结果表明，Faster-RCNN 作为 two-stage 目标检测的代表算法，其 mAP 值大于 SSD 和 YOLO 这些 one-stage 算法，但是检测速度和检测精度呈现出一种不平衡的状态，其速率仅有 20.4 fps，在检测速度上远远低于 one-stage 类的检测算法。相较于其他 one-stage 算法，YOLOv4 算法已经取得较好的检测效果，检测精度和速度也较为平衡。而本文提出的 YOLOv4-ST 算法的检测速度和精度均优于其他检测模型。

图 7 为本文的 YOLOv4-ST 模型与其他模型的检测效果对比图。由于受到摄像头视距及位置影响，学生行为检测属于小目标检测，且目标较多，因此本文改进了 YOLOv4 网络模型的特征提取结构和尺度预测阶段，以提高模型检测能力。从检测效果对比图中可以看出，YOLOv4-ST 在对学生课堂异常行为进行检测时，模型的错检漏检情况相较于 YOLOv3、YOLOv4 网络模型有了明显改善，对于检测目标的类别置信度也有一定提高，较好地实现了多个小目标行为的检测。

4 结论

本文在 YOLOv4 目标检测算法的基础上，提出了一种改进 YOLOv4 算法的学生课堂行为检测方法。借鉴 MobileNet 思想，设计了 MoblieDarknet19 特征提取结构，降低了网络参数和计算量；此外，使用新的“梯”形特征融合方式，得到梯形-MoblieDarknet19，在避免过多增加网络参数和计算量的情况下，有效提升了目标特征信息传输能力和网络学习效果；在预测阶段，引入了密集连接结构，通过多层特征融合的方式，提高网络对小目标的检测精度。经过实验对比，本文提出的 YOLOv4-ST 网络对学生课堂行为检测有一定的适用价值和研究意义。之后本文将继续研究学生课堂行为的检测改进方法，制作包含数量更多、学生行为类型更为全面的数据集，针对错检漏检问题，后续会尝试选取更加有效的关键区域进行特征学习和检测。

参考文献

[1] PAK M，KIM S.A review of deep learning in image recognition[C]//International Conference on Computer Applications and Information Processing Technology(CAIPT)，2017.

[2] 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 & Machine Intelligence，2017，39(6)：1137-1149.

[3] LIU W，ANGUELOV D，ERHAN D，et al.SSD：single shot multibox detector[C]//Proceedings of European Conference on Computer Vision.Berlin，Germany：Springer，2016：21-37.

[4] REDMON J，DIVVALA S，GIRSHICK R，et al.You only look once：unified，real-time object detection[C]//IEEE Conference on Computer Vision and Pattern Recognition，LasVegas.NV.IEEE，2016：779-788.

[5] REDMON J，Farhadi A.YOLO9000：bettrt，faster，stronger[C]//IEEE Conference on Computer Vision and Pattern Reconganiton，Honolulu，HI.IEEE，2017：6517-6525.

[6] REDMON J，FARHADI A.YOLOv3：an incremental improvement[J].Computer Vision and Pattern Recognition，2018，36(5)：1658-1672.

(a)YOLOv3

(b)YOLOv4

(c)对比文献[20]

(d)YOLOv4-ST

图 7　YOLOv4 与 YOLOv4-ST 模型检测效果对比

[7] ZHENG R, JIANG F, SHEN R M. Intelligent student behavior analysis system for real classrooms[C]//IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020.

[8] LIU D, JI Y, YE M, et al. An improved attention-based spatio-temporal-stream model for action recognition in videos[J]. IEEE Access, 2020, 8(8): 61462-61470.

[9] BOCHKOVSKIY A, WANG C, LIAO H. YOLOv4: optimal speed and accuracy of object detection[J]. arXiv: 2004.10934, 2020.

[10] WANG C, LIAO H, WU Y, et al. CSPNet: a new backbone that can enhance learning capability of CNN[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops(CVPRW), 2020: 1571-1580.

[11] REN X, YANG D. Student behavior detection based on YOLOv4-Bi[C]//2021 IEEE International Conference on Computer Science, Artificial Intelligence and Electronic Engineering(CSAIEE), 2021: 288-291.

[12] MISRA D. Mish: a self regularized non-monotonic neural activation function[J]. arXiv: 1908.08681, 2020.

[13] HE K, ZHANG X, REN S, 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] LIU S, QI L, QIN H, et al. Path aggregation network for instance segmentation[C]//2018 IEEE Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 8759-8768.

[15] 李彬, 汪诚, 丁相玉, 等. 改进 YOLOv4 的表面缺陷检测算法[J/OL]. 北京航空航天大学报. 1-10[2021-12-24]. https://doi.org/10.13700/j.bh.1001-5965.2021.0301.

[16] LIN T Y, DOLLAR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 936-944.

[17] LI H P, LI C Y, LI G B, et al. A real-time table grape detection method based on improved YOLOv4-tiny network in complex background[J]. Biosystems Engineering, 2021, 212: 347-359.

[18] SANDLER M, HOWARD A, ZHU M, et al. MobileNetV2: inverted residuals and linear bottlenecks[C]//Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition(CVPR), Salt Lake City, UT, USA, 2018.

[19] HUANG G, LIU Z, LAURENS V, et al. Densely connected convolutional networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Honolulu, Hawaii, USA. IEEE Computer Society, 2017: 4700-4708.

[20] DENG H F, CHENG J H, LIU T, et al. Research on iron surface crack detection algorithm based on improved YOLOv4 network[J]. Journal of Physics: Conference Series, 2020, 1631(1): 012081.

(收稿日期: 2022-04-10)

作者简介:

张颖(1982-), 女, 博士, 讲师, 主要研究方向: 深度学习、通信。

张喆(1997-), 通信作者, 女, 硕士研究生, 主要研究方向: 深度学习、目标检测, E-mail: fightingzhe@163.com。

龙光利(1968-), 男, 硕士, 教授, 主要研究方向: 电子技术应用、物联网。

扫码下载电子文档

版权声明