计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2020,29(1):203−208 [doi: 10.15888/j.cnki.csa.007240]

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

基于 DE-YOLO 的室内人员检测方法①

张明伟[1], 蔡坚勇[1,2,3,4], 李　科[1], 程　玉[1], 曾远强[1]

[1](福建师范大学 光电与信息工程学院, 福州 350007)
[2](福建师范大学 医学光电科学与技术教育部重点实验室, 福州 350007)
[3](福建师范大学 福建省光子技术重点实验室, 福州 350007)
[4](福建师范大学 福建省光电传感应用工程技术研究中心, 福州 350007)
通讯作者: 蔡坚勇, E-mail: cjy@fjnu.edu.cn

摘　要: 目标检测的一个重要应用场景是对室内流动人员的检测与定位, 为了降低模型的冗余度和提高检测的精确度, 因此本文提出一种基于 DE-YOLO 的室内人员检测方法. 通过使用 K-means 算法对数据集进行聚类, 并设计出这种 DE-YOLO 深度卷积神经网络结构. 通过 DE-YOLO 网络结构中的密集型连接, 实现模型大小的压缩和特征信息的复用, 最后对提取到的特征进行目标检测. 在 VOC2012 数据集上进行实验表明, 新改进的深度卷积网络应用性能有较大的提升.

关键词: 深度学习; 目标检测; YOLO v3; K-means; 室内人员检测

引用格式: 张明伟,蔡坚勇,李科,程玉,曾远强.基于 DE-YOLO 的室内人员检测方法.计算机系统应用,2020,29(1):203–208. http://www.c-s-a.org.cn/1003-3254/7240.html

Indoor Personnels Detection Method Based on DE-YOLO

ZHANG Ming-Wei[1], CAI Jian-Yong[1,2,3,4], LI Ke[1], CHENG Yu[1], ZENG Yuan-Qiang[1]

[1](College of Photonic and Electronic Engineering, Fujian Normal University, Fuzhou 350007, China)
[2](Key Laboratory of Optoelectronic Science and Technology for Medicine of Ministry of Education, Fujian Normal University, Fuzhou 350007, China)
[3](Fujian Provincial Key Laboratory of Photonics Technology, Fujian Normal University, Fuzhou 350007, China)
[4](Fujian Provincial Engineering Research Center for Optoelectronic Sensors and Intelligent Information, Fuzhou 350007, China)

Abstract: An important application scenario for target detection is the detection and location of indoor mobile personnel. In this study, we propose an indoor personnel detection method to improve YOLOv3. First, the proposed method clusters the dataset by using K-means algorithm and designs a DE-YOLO deep convolutional neural network structure. Through the dense connection in the DE-YOLO network structure, the compression of the model sizes and the reuse of the feature information are realized. Finally, the target detection is performed on the extracted features. Experiments show that the application of the newly improved deep convolutional network has greatly improved application effect on VOC2012 datasets.

Key words: deep learning; target detection and localization; YOLO v3; K-means; indoor personnels detection

1　引言

目标检测的场景分为室内和室外, 室内环境的变化虽然不如室外环境那么复杂, 但它们对于运动物体的检测也将产生显著的影响. 由于对室内人员检测的需求性更强, 本文主要研究室内人员的检测. 比如, 教室学生检测就是室内场景下人员检测的一个重要的方

① 基金项目: 福建省自然科学基金 (2017J01744)
Foundation item: Natural Science Foundation of Fujian Province (2017J01744)
收稿时间: 2019-06-27; 修改时间: 2019-07-16; 采用时间: 2019-07-19; csa 在线出版时间: 2019-12-27

向. 针对教育中的室内人员检测问题, 本文完全可以通过计算机视觉的相关技术——目标检测, 从而解决教室学生检测问题.

有关目标检测的技术, 分为传统算法和深度学习算法. 传统算法主要分为目标实例检测与传统目标类别检测. 自 2010 年, 深度学习成为计算机视觉的主要研究方向, 使用卷积神经网络进而大幅提高了图像检测的准确率, 因此越来越多的人将深度学习的思想应用到目标检测类别检测中. 在这方面, 基于深度学习的目标检测与识别算法已经成为主流, 主要有三大类: 基于快速 CNN 的目标检测技术, 如 R-CNN、Mask RCNN[1]等; 基于回归学习的目标检测与识别, 如 SSD[2]、YOLO 等; 基于学习搜索的目标检测与识别, 如 AttentionNet、FSRL[3]等. 其中 YOLO 系列的算法是一个端对端的模型, 其模型结构复杂度要优于 R-CNN 系列, 很适合对实时性要求较高的应用场景[4].

本文采用回归的目标检测与识别方法, 以深度学习网络 YOLO v3 为基础, 将教室中的学生作为待检测目标. 因为检测目标只有室内人员, 为了降低模型的冗余度和提高检测的精确度, 提出一种 DE-YOLO 神经网络结构, 对网络的结构进行改进, 并对层级结构中的参数进行调整, 使得不仅减少模型的占用空间大小, 而且能准确识别教室中的学生. 通过本文对 DE-YOLO 和 YOLO v3 的实验结果对比, DE-YOLO 运行速度明显优于 YOLO v3, 同时保证了预测准确率. 并且基于内存大小为 8 GB 和型号为 Inter i5 的 CPU 硬件环境处理, 不使用 GPU 情况下, 检测速度提升了 3FPS.

2 相关工作

2.1 YOLO v3 的原理

2016 年 Joseph Redmond 等人提出了 YOLO (You Only Look Once) 算法, 它主要基于回归学习, 实现用单一网络对图片只要看一次就能检测与识别目标[5]. 通过完善发展, 于 2018 年提出 YOLO v3, 也是目前效率最高的版本[6]. YOLO v3 依然保持了 YOLO v2 的快速检测, 并大大提高了识别的正确率, 尤其是在小目标的检测与识别上, 识别率也有较大的提升. 相对于 YOLO v2, YOLO v3 结合 ResNet 的思想, 运用了若干个 ResNet 模块[6]. YOLO v3 在网络框架方面, 大量使用具有良好表征能力的 3×3 和 1×1 的卷积层, 并网络结构中不断穿插着一些 ResNet. 最终 YOLO v3 整体的网络结构中包含了 53 个卷积层, 因此 Joseph Redmond 也把它称为 Darknet-53, 如图 1 所示.

	Type	Filters	Size	Output
	Convolutional	32	3×3	256×256
	Convolutional	64	3×3/2	128×128
1×	Convolutional	32	1×1	
	Convolutional	64	3×3	
	Residual			128×128
	Convolutional	128	3×3/2	64×64
2×	Convolutional	64	1×1	
	Convolutional	128	3×3	
	Residual			64×64
	Convolutional	256	3×3/2	32×32
8×	Convolutional	128	1×1	
	Convolutional	256	3×3	
	Residual			32×32
	Convolutional	512	3×3/2	16×16
8×	Convolutional	256	1×1	
	Convolutional	512	3×3	
	Residual			16×16
	Convolutional	512	3×3/2	8×8
4×	Convolutional	512	1×1	
	Convolutional	1024	3×3	
	Residual			8×8
	Avgpool		Global	
	Connected		1000	
	Softmx			

图 1 Darknet-53 网络结构图

借鉴 Faster RCNN 的思想, YOLO v3 还引入了多尺度预测方式. 每种尺度预测 3 个 boxes, anchor 的设计方式依然使用聚类, 将其按照大小均分给 3 种尺度[7]. 同时网络结构中最后的分类器也从 Softmax 函数改为 logistic 函数, 使得能够支持多种类型目标的检测与定位.

2.2 DenseNet 的原理

2017年Huang等提出了DenseNet (Densely connected convolutional Networks) 网络[8], 主要对 ResNet 和 Inception 两种网络的对比学习: 如果卷积神经网络在每个单元的输入及输出之间有更短的连接, 它实质上有更深、更精准、训练更高效的特点. DenseNet 的本质是在于对目标特征的学习, 通过的表征信息的最大化利用, 来达到网络模型的最简化和最优化, 尽可能降低参数冗余.

DenseNet 网络结构中内嵌 3 个 dense block, 每个 dense block 中串连着 4 个卷积层, 在每个 dense block 中, 可以把每个卷积层之前所有前置卷积层的输出汇总为输入. 每个 dense block 的结构如图 2 所示, 层与层之间可以用池化层相连.

DenseNet 引入这样的 dense block 有如下优点:

(1) 由于网络中每层都会接受前面所有层的特征输入, 为了避免随着网络层数的增加, 后面层的特征维

度增长过快, 在每个阶段之后进行下采样时, 会首先通过一个卷积层将特征维度压缩至当前输入的一半, 再进行池化操作, 即解决梯度消失的问题;

(2) 通过每层之间的跳跃连接, 加强了网络模块之间的信息交流, 其本质就是特征的复用. 与 ResNet 不同的是, 这样的密集型连接, 使得信息流更大, 不是简单的叠加效果, 使得小模型产生大数据.

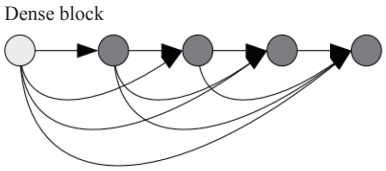

图 2 密集卷积块结构图

本文正是对 DenseNet 的上述优点的充分考虑, 在 YOLO v3 上提出 DE-YOLO 网络, 达到对目标特征的重复的学习和利用, 使得算法在室内人员检测方面有着更优的效果.

3 DE-YOLO 的设计

本文考虑到室内人员检测时, 存在检测目标比较密集, 且先验框大小不一. 比如, 教室中学生的检测就是一个很好的例子. 所以, 首先需要对数据集进行预处理, 再对网络层级结构进行删减和替换, 并将密集连接的思想更新进去, 搭建一个全新的网络, 最后达到设计的效果.

3.1 基于 K-means 的数据集预处理

当神经网络来检测一幅图像中的多个目标时, 其实网络实际上需要大量的先验框执行预测, 并只显示出它确定为一个对象的那些检测结果. 由于 R-CNN 系列中先验框的高度和宽度都是手动设置的, 客观性较差. 如果选取的的先验框的高度和宽度比较合适, 所得的模型的性能将更优, 使得预测效果更好. 所以, YOLO v3 中为了针对数据集的目标框大小进行聚类分析, 可采用 K-means 算法.

K-means 算法是一种经典的聚类算法, 通常使用欧几里得度量等方式作为两个样本相似程度的评价指标[9]. 因为先验框设置的最初目的是为了使得 ground truth 与预测框的重合度尽可能高, 即式 (1) 中的交并比 *IOU* 的值越接近为 1 越好. 但是由于这些经验值不一定适用于对室内人员检测的场景, 会对最终的检测产生一定的干扰. 例如使用欧几里得度量会让大的边界区域比小的边界区域更易出现误差, 导致精确度下降. 我们希望通过先验框来获取良好的 *IOU*, 因为它不受边界框的尺寸影响. 因此我们选择新的方式来表达 *IOU*, 如下[10]:

$$d(box, centre) = 1 - IOU(box, centre) \tag{1}$$

其中, *centre* 表示为类簇中心, *box* 表示为目标, *IOU* (*box*, *centre*) 表示类簇中心框和目标框的交并比. 交并比 *IOU* 表示预测框的准确程度, 其公式为:

$$IOU(bb_{gt}, bb_{dt}) = \frac{bb_{gt} \cap bb_{dt}}{bb_{gt} \cup bb_{dt}} \tag{2}$$

其中, bb_{gt} 表示真实框, bb_{dt} 表示预测框.

由于 K-means 算法具有收敛于局部最优解的特性, 所以本文起初会选取多组初始值, 对其分别运行算法, 如果获得目标函数值最小, 则选取这一组方案作为最后聚类结果. 最终的聚类结果受初始化的影响很大, 一般采用随机的方式生成中心点[11], 对于比较有特点的数据集可采用一些启发式的方法选取中心点. 如果目标的边界框大小太多, 反而会增加一定的计算量而导致效率降低. 所以, 实验选取 *K*=[1, 20], 统计出不同锚点框数量 (*K* 的大小) 下所对应的 *IOU* 值, 具体的关系如图 3 所示.

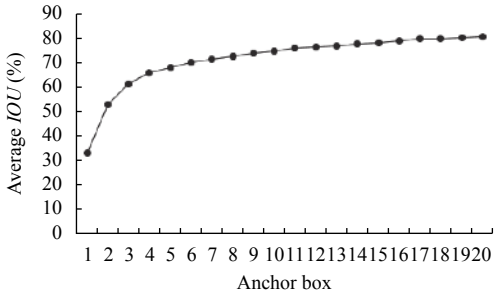

图 3 锚点框数量与平均交并比的关系

根据图中锚点框数量与平均交并比的关系可知, 在 *K*=6 之后, 曲线变化趋于稳定, 且先验框数量合理, 不会带来过多的计算开销, 所以本文得到的 6 个聚类的中心为 (10, 14)、(23, 27)、(37, 58)、(81, 92)、(136, 169)、(344, 319).

根据 Joseph Redmond 的工作, YOLO v3 在 COCO 和 VOC 数据集上分在 32×32、16×16、8×8 这 3 个不同的尺度上进行预测[12], 最后判断最终结果, 如图 4 所示.

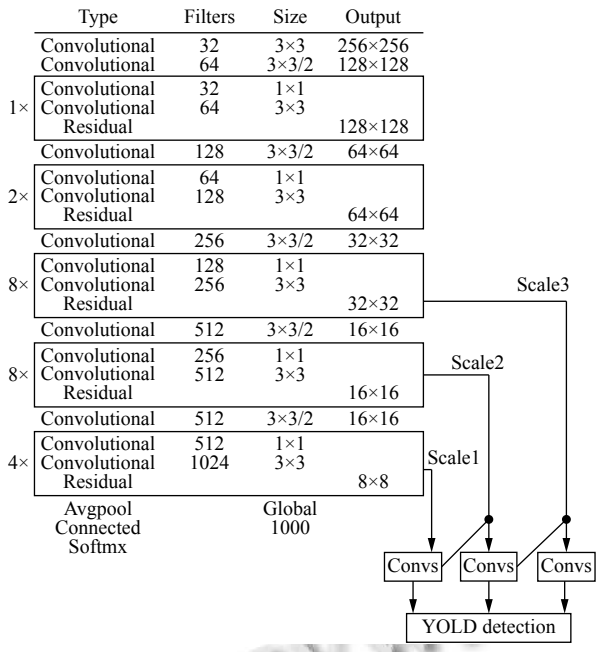

图 4 YOLO v3 多尺度预测方式

图 4 中 YOLO v3 分别在 3 个尺度上对目标进行预测: 尺度 1 上, 在基础网络之后添加一些卷积层再输出 box 信息; 尺度二上, 在尺度 1 进行上采样再与 16×16 大小的特征图相加, 卷积输出 box 信息; 尺度 3 上, 与尺度 2 类似, 使用了 32×32 大小的特征图.

3.2 DE-YOLO 网络结构的设计

随着网络层数的加深, 虽然 ResNet 模块可以缓解梯度爆炸的现象, 使得精确度不会随之降低. 但是, 这是基于网络结构可以比较复杂, 对内存占用没有较高要求的前提下. 为了尽可能减少网络结构的复杂度, 降低网络模型对内存的占用, 并保证较高的精确度. 使用 Dense block 模块将表现的比 ResNet 模块更好.

本文借鉴 DenseNet 网络的思想, 为了压缩模型并提高特征信息的复用率, 需要对网络结构进行调整. 考虑到在 32×32、16×16 尺度的特征图上, 包含较多的表征信息, 而在 8×8 尺度上的表征学习能力有限. 所以将 YOLO v3 这三个尺度上的 ResNet 模块替换为与其维度相适应的 Dense block 模块, 其更新的网络如图 5 所示. 通过对尺度 2、尺度 3 构建这样一种密集连接的网络结构, 使得不同维度学习到的表征信息得以极致的利用和汇总, 为下一步的精准预测提供的有效的保证.

4 实验分析

4.1 实验环境与数据

实验环境如表 1 所示.

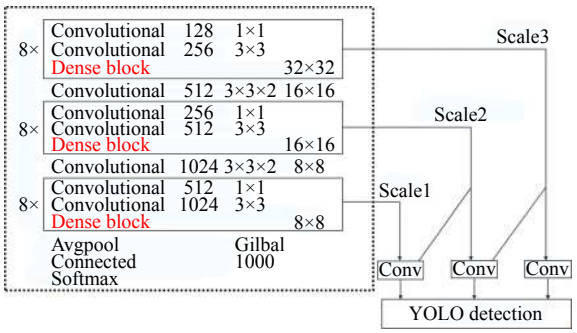

图 5 DE-YOLO 网络结构

表 1 实验环境

硬件与软件平台	
操作系统	Win10
CPU	Inter(R)Core(TM)i5-4460 @3.20 GHz
内存	4 GB
硬盘	500 GB
开发平台	Anaconda3+keras
开发语言	Python3.6

为了保证实验结果的可靠性, 数据集的选用十分重要. Pascal Visual Object Calsses (VOC)[13]是计算机视觉领域中的一个公认的数据集, 具有一定权威性. 本文选用 VOC2012 数据集, 并提取了数据集中 1000 张不同的 person 照片. 随机抽取 820 张图片作为训练集, 80 张图片作为验证集, 100 张图片作为测试集.

4.2 实验设计

数据集用 labelImg 软件标注完成后, 分别对 YOLO v3 和 DE-YOLO 网络进行训练. 实验过程中, 网络的学习率 (learning rate) 为 10^{-4}, 梯度下降的优化器选用 Adma, 以便快速收敛并正确学习, 训练迭代次数为 1000. 实际训练过程中, 为了避免过拟合情况的出现, 每迭代 50 次对模型进行保存, 输出后缀为.h5 文件.

本文算法流程图大致如图 6 所示.

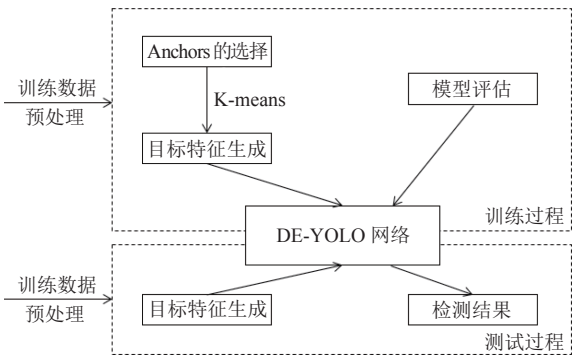

图 6 算法流程图

从图中可以看出, 实验分为训练和测试两个部分, 分别对训练数据和测试数据进行预处理. 通过对训练集的 K-means 聚类分析, 得到相应目标特征, 并通过 DE-YOLO 模型进行训练. 最后进行测试并模型评估.

本文方法总体流程图, 如图 7 所示.

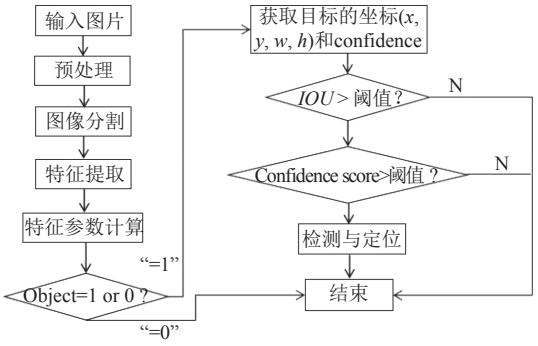

图 7 室内人员检测方法流程图

4.3 模型评估

为了避免数据集中的常出现的类不平衡的现象, 使得无论正负样本如何变化, 都不影响模型表达的准确性. 本文采用准确率 P (precision)、召回率 R (recall)、精确率 ACC (accuracy)、F_1 值 (F_1-score) 值作为评价指标, 通过 ROC 曲线图来评估其模型的性能[14].

precision表示被分为正例的示例中实际为正例的比例, 其中设 TP 为将正类预测为正类数, FP 为将负类预测为正类数, 公式如下:

$$P = TP/(TP+FP) \tag{3}$$

recall是覆盖率的度量, 度量有多少个正例被分为正例, 其中设 FN 为将正类预测为负类数, 公式如下:

$$R = TP/(TP+FN) \tag{4}$$

accuracy是被分类正确的样本数占总样本的比例, 其中设 TN 为将负类预测为负类数, 公式如下:

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \tag{5}$$

F_1 为precision和recall的加权平均调和, 公式如下:

$$F_1 = \frac{2\times P\times R}{P+R} \tag{6}$$

4.4 实验结果

为了评估 DE-YOLO 性能, 与 YOLO v3 进行对比, 其迭代次数与精确率关系对比如表 2 所示. DE-YOLO 算法较于 YOLO v3 算法更能有效的对室内的人员进行检测.

表 2 迭代次数与精确率关系对比 (单位: %)

迭代次数	YOLO v3	DE-YOLO
…	…	…
400	67.26	75.77
600	74.89	83.22
800	86.15	**93.01**
1000	**94.53**	96.78
…	…	…

从表 2 中可以看出, DE-YOLO 在迭代 800 次的精确率已达到 93.01%, 与 YOLO v3 迭代 1000 次时的精确率相差无几. 当 DE-YOLO 迭代 1000 次时, 比 YOLO v3 的精确率提高了 2.38%. 对于 DE-YOLO 网络误检情况, 选取部分经典的实验结果, 如图 8 和图 9 所示.

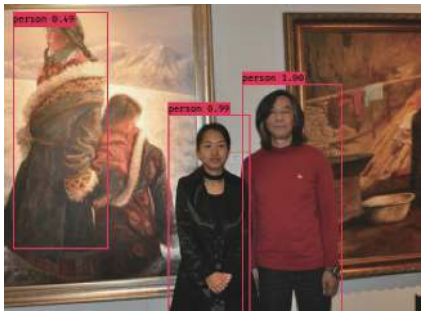

图 8 DE-YOLO 误检情况 (a)

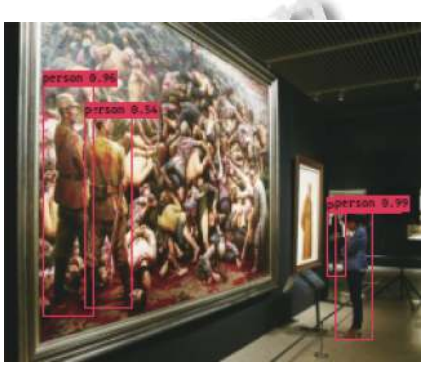

图 9 DE-YOLO 误检情况 (b)

在图 8 和图 9 中存在的误检情况, 主要是由于神经网络很难区分现实与画像中的人物, 只要是是符合目标的特征的对象, 都将被检测输出, 很难避免.

绘制 YOLO v3 和 DE-YOLO 的 ROC 曲线, 如图 10 所示. 这里引入 AUC (Area Under roc Curve) 概念, 即 ROC 曲线下的面积大小[15]. AUC 的值越接近为 1, 则模型性能越突出. 由图 10 可得, 本文的 DE-YOLO 模型性能优于 YOLO v3 的原网络.

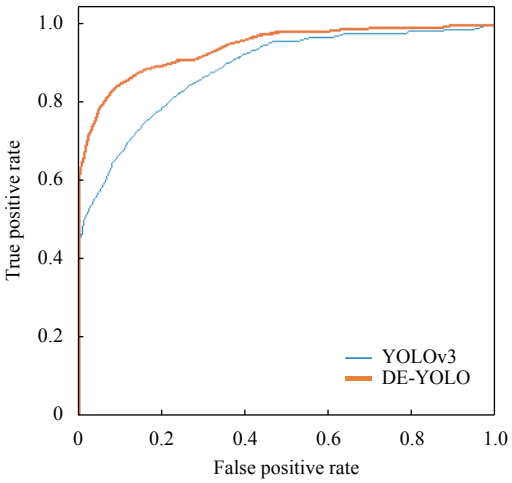

图 10 ROC 曲线图

5 结语

本文提出了一种基于密集型连接的 DE-YOLO 卷积神经网络, 旨在通过网络之间的密集型, 连接提高其通道中的特征信息的利用率, 并减少内存占用空间. 另外, 网络通过 K-means 对数据集的预处理, 提高对室内人员检测的精确性. 通过实验表明, DE-YOLO 在保证与 YOLO v3 相近正确率的情况下, 减少了模型大小和内存占用空间, 可以将模型大小从 235 MB 减少至 33 MB, 实现了轻量化处理. 另外, 由于存在数据较少、目标标注引入干扰背景的问题, DE-YOLO 检测的精确度提升会遇到瓶颈, 同时网络结构如何进一步的压缩和裁剪也是一个值得研究的方向, 后期的工作将针对这些问题进入深入的研究.

参考文献

1 Ren SQ, He KM, Girshick R, *et al*. Faster R-CNN: Towards real-time object detection with region proposal networks. arXiv: 1506.01497, 2016.

2 Liu W, Anguelov D, Erhan D, *et al*. SSD: Single shot MultiBox detector. Proceedings of the 14th European Conference on Computer Vision. Amsterdam, The Netherlands. 2016. 21–37.

3 焦李成, 赵进, 杨淑媛, 等. 深度学习、优化与识别. 北京: 清华大学出版社, 2017. 341–342.

4 Redmon J, Farhadi A. YOLO9000: Better, faster, stronger. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. 2016. 6517–6525.

5 Redmon J, Divvala S, Girshick R, *et al*. You only look once: Unified, real-time object detection. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2016. 779–788.

6 Redmon J, Farhadi A. YOLOv3: An incremental improvement. arXiv: 1804.02767, 2018.

7 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2016. 770–778.

8 Huang G, Liu Z, Vab Der Maaten L, *et al*. Densely connected convolutional networks. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. 2017. 2261–2269.

9 Coates A, Ng AY. Learning feature representations with K-means. in: Montavon G, Orr GB, Müller KR, eds. Neural Networks: Tricks of the Trade. 2nd ed. Berlin, Heidelberg: Springer, 2012. 561–580.

10 郑志强, 刘妍妍, 潘长城, 等. 改进 YOLO v3 遥感图像飞机识别应用. 电光与控制, 2019, 26(4): 28–32.

11 魏杰. 基于 K-means 聚类算法改进算法的研究. 信息与通信, 2018, (5): 14–15.

12 Girshick R, Donahue J, Darrell T, *et al*. Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA. 2014. 580–583.

13 PASCAL VOC Challenge performance evaluation and downloadserver.http://host.robots.ox.ac.uk:8080/leaderboard/main_bootstrap.php.

14 陈颖熙, 廖晓东, 苏例月, 等. 基于 GDBN 网络的文本情感倾向分类算法. 计算机系统应用, 2019, 28(1): 163–168. [doi: 10.15888/j.cnki.csa.006723]

15 何长婷. 课堂签到系统中的人脸识别方法研究与实现[硕士学位论文]. 合肥: 中国科学技术大学, 2018. 46–48.