Hans Journal of Data Mining 数据挖掘, 2023, 13(1), 10-22
Published Online January 2023 in Hans. https://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2023.131002

Hans汉斯

基于改进YOLO V3算法的室内人数统计模型

何　强[1,2*]，李　静[1,2]，陈琳琳[1,2]

[1]北京建筑大学理学院，北京
[2]北京建筑大学，大数据建模理论与技术研究所，北京

收稿日期：2022年11月23日；录用日期：2022年12月23日；发布日期：2023年1月3日

摘　要

基于机器学习与深度学习的目标检测方法被广泛应用于人数统计，然而实际检测区域往往存在人群相互遮挡，或光照不均匀等情况时，人数统计仍然面临很大挑战。为此，提出了一种改进的YOLO V3模型，使其更好的适应室内人群的人数统计。首先自建并丰富了数据集，增加了训练数据的多样性，并通过K-means算法重新聚类锚框；其次，提出了F-YOLO V3模型，该模型中增加104 × 104尺寸的特征图输出并取消13 × 13尺寸特征图的输出；将原网络每一层上采样后的特征图继续上采样，得到的特征图与原网络相应尺寸的特征图进行拼接；并将输出层前的5个卷积变成了1个卷积和2个残差单元，提取更多特征信息，增强对模糊或者较小目标检测能力；最后增加一个ADIOU Loss分支衡量检测框的定位准确度。实验结果表明，F-YOLO V3模型具有更高的召回率和平均精度，室内场景下的人员统计性能得到显著提升。

关键词

目标检测，YOLO V3，特征提取网络，多尺寸检测算法，ADIOU Loss

Indoor People Counting Model Based on Improved YOLO V3 Algorithm

Qiang He[1,2*], Jing Li[1,2], Linlin Chen[1,2]

[1]School of Science, Beijing University of Civil Engineering and Architecture, Beijing
[2]Institute of Big Data Modelling and Technology, Beijing University of Civil Engineering and Architecture, Beijing

Received: Nov. 23rd, 2022; accepted: Dec. 23rd, 2022; published: Jan. 3rd, 2023

Abstract

Object detection methods based on machine learning and deep learning model are widely used in

*通讯作者。

文章引用: 何强, 李静, 陈琳琳. 基于改进 YOLO V3 算法的室内人数统计模型[J]. 数据挖掘, 2023, 13(1): 10-22.
DOI: 10.12677/hjdm.2023.131002

people counting. However, when there are too many objects in the same area, people will be occluded, or people in the video are not easy to find in the dark, people counting is still a big challenge. Therefore, an improved YOLO V3 model is proposed to better adapt to the number of indoor crowd statistics in classrooms. Firstly, the data set was self-built and enriched to increase the diversity of training data, and the anchor boxes were re-clustered by K-means algorithm. Secondly, the YOLO V3 feature extraction network and multi-dimension detection algorithm were improved, and the F-YOLO V3 model was proposed. In this model, the output of 104 × 104 feature map was added and the output of 13 × 13 feature map was canceled. The sampled feature images of each layer of the original network are continued to be sampled, and the obtained feature images are spliced with the corresponding size feature images of the original network. The 5 convolutions in front of the output layer are changed into 1 convolution and 2 residual units to extract more feature information and enhance the detection ability of fuzzy or small targets. Add an ADIOU Loss branch to measure the positioning accuracy of the detection box; Finally, the real-time number of people in the output screen is counted. The experimental results show that the improved YOLO V3 algorithm has higher recall rate and average precision, and the performance of personnel statistics in indoor scenes is significantly improved.

Keywords

Object Detection, YOLO V3, Feature Extraction Network, Multi Size Detection Algorithm, ADIOU Loss

1. 引言

目标检测是目前计算机视觉和数字图像处理领域的一个重要研究方向，在目标检测的实际应用场景中，室内人数统计一直以来都是一个具有挑战性的问题。像教室、商超、银行及候车室等地方，都需要进行人数的统计，这些场所往往人数较多，位置分散，且监控画面往往是俯拍角度，所以视频中的人像较小的同时也存在相互遮挡问题，因此如何有效的进行特征提取，提高室内人数统计的准确度不仅具有理论意义，也具有重要的实用价值。

目前基于深度学习[1] [2] [3] [4]且应用前景[5] [6]比较广泛的目标检测[7] [8] [9]算法可以分为两类：1) 两阶段目标检测算法：基于 Region Proposal 的 R-CNN (Region-Convolutional Neural Network) [10]系列算法，需要先产生目标候选框，卷积神经网络对候选框做分类与回归。常见的算法有 Fast R-CNN [11]、SPP-Net [12]、Corner-Net [13]、Faster R-CNN [14]、Mask R-CNN [15]等。2) 一阶段目标检测算法：不需要产生候选框，仅使用卷积神经网络直接将目标框定位问题转化为回归问题，预测不同目标的类别与位置。常见的算法有 YOLO (You Only Look Once)系列[16] [17] [18]、SSD (Single Shot MultiBox Detector) [19]等。基于候选框的两阶段方法，经过两次分类和位置回归，在检测准确率和定位精度上占优。由于存在选框和检测两个阶段，会比一阶段算法效率低，而室内人数统计往往需要满足实时性，因此一阶段目标检测算法更适合用来进行人数统计。

近年来，用目标检测进行人数统计的研究取得了一定的突破，陈晓[20]等人针对目标检测中的误检漏检问题，提出了一种基于视频的人数统计方法，通过对特征提取、损失函数以及后处理阶段的改进，使得检测准确率以及召回率有一定提高，且处理速度较快。成玉荣[21]等人为了统计当前监控环境下的人数，

引入了通道注意力机制，改进了 Tiny-YOLOv3 算法，训练人体头部的目标检测模型，实验的平均检测精度高达 80%。郑国书[22]等人基于 SSD 模型，提出了一种基于人头检测的视频室内人数统计方法，该算法可以对小尺度人头进行检测，准确度高，实时性好，但由于 SSD 模型使用低级特征检测小目标，特征提取不够充分。

基于 YOLO V3 检测算法，本文提出了一种改进的 YOLO V3 室内人数统计模型，并应用于教室人数识别统计中。在自建室内人群检测数据集中训练新模型，结果表明改进后的 YOLO V3 算法能更好的提取特征，测试的召回率和平均精确度有明显提升。本文的主要贡献如下：

1) 自建数据集包括动态视频数据和静态图片数据，且利用对比度增强、亮度增强等方法增加数据量，提高了训练数据的多样性；

2) K-means 算法重新聚类锚框；

3) 提出了 F-YOLO V3 模型，其中改进了 YOLO V3 特征提取网络以及多尺寸检测算法。改进点包括：利用低层特征图包含更多特征细节的特点，增加 104 × 104 尺寸的特征图输出，进行类别判断与边框预测；为提高网络对小目标的召回率和检测准确度，取消了 13 × 13 尺寸特征图的输出；用上采样小特征图后与大特征图拼接的方式融合不同粒度的特征图，进而从小目标中得到更细粒度的特征以及位置信息；将输出层前的 5 个卷积变成了 1 个卷积和 2 个残差单元，以此减少人头小目标在复杂场景的漏检率，提高网络对小目标的检测率；

4) 增加一个 ADIOU (Area Distance) loss 分支，增强检测框定位准确度。

2. 相关知识

2.1. YOLO V3 模型

YOLO V3 算法以 YOLO V1 和 YOLO V2 算法为基础。YOLO V2 借鉴了 R-CNN 的思路，引入了锚框(anchor)，并通过聚类进行选取，增加了细粒度特征，将浅层特征图连接到深层特征图，网络修改为全卷积网络，YOLO V3 进一步加入了特征金字塔网络的思想，利用多尺度特征进行对象检测，在保持速度优势的前提下，提升预测精度，尤其是加强了对小目标的识别能力。

在基本的图像特征提取方面，YOLO V3 采用了称之为 Darknet-53 的网络结构，共包含 53 层卷积层，Darknet-53 由 5 个残差模块构成，每个残差模块由多个残差单元[23]组成。每个残差单元由两个 CBL 单元和一个快捷链路构成，其中 CBL 单元包含卷积层、Batch Normalization 层和 Leaky Relu (Rectified Linear Unit) [24]激活函数，这样有利于解决深层次网络的梯度问题。YOLO V3 网络结构如图 1 所示。

2.2. YOLO V3 损失函数

YOLO V3 的损失函数由四项组成，分别是预测框的中心点坐标损失、预测框的宽高损失、置信度损失和类别损失，其损失函数计算式见式(1)。

$$\text{Loss} = L_1 + L_2 + L_3 + L_4 \tag{1}$$

其中：L_1 是预测框的中心点坐标损失，其详细公式如式(2)：

$$L_1 = \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{ij}^{obj} \left[\left(x_i - \hat{x}_i^j \right)^2 + \left(y_i - \hat{y}_i^j \right)^2 \right] \tag{2}$$

x_i、y_i 分别是预测第 i 个网格的第 j 个锚框的中心点坐标，与之对应的真实值为 $\hat{x}_i^j$、$\hat{y}_i^j$。λ_{coord} 为权重系数，I_{ij}^{obj} 表示为分段函数：当第 i 个网络的第 j 个锚框包含检测目标对象的归一化值时 $I_{ij}^{obj} = 0$，否则 $I_{ij}^{obj} = 1$。

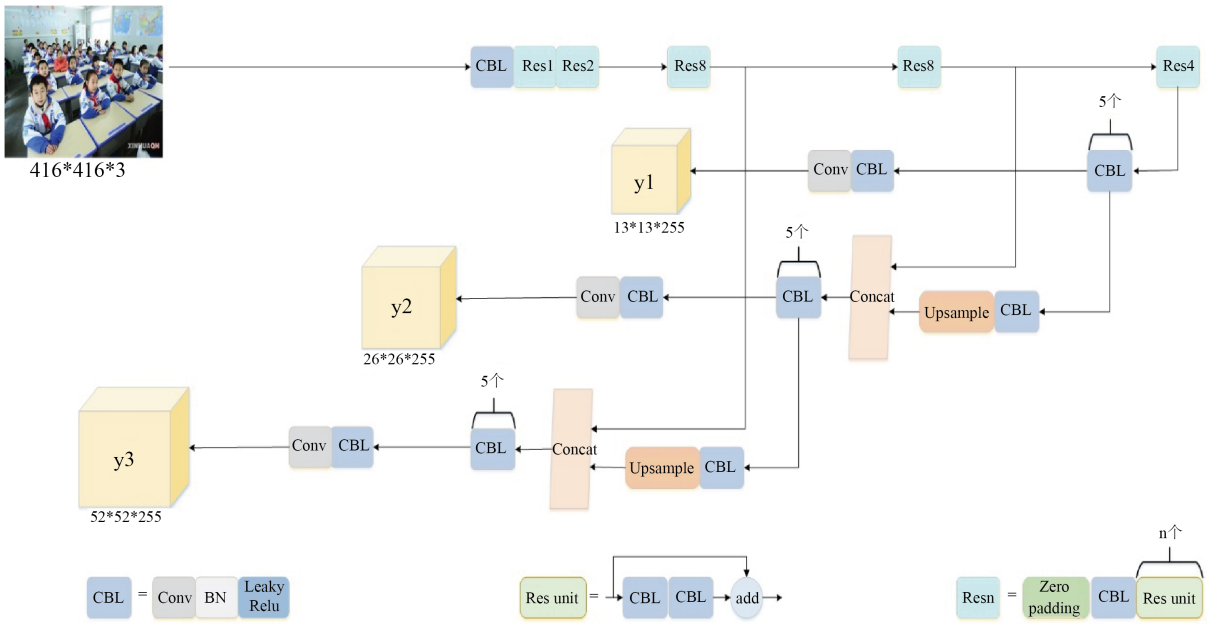


Figure 1. YOLO V3 network structure
图 1. YOLO V3 网络结构

L_2 是预测框的宽高损失，其详细公式如式(3)：

$$L_2 = \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{ij}^{obj} \left[\left(\sqrt{w_i^j} - \sqrt{\hat{w}_i^j} \right)^2 + \left(\sqrt{h_i^j} - \sqrt{\hat{h}_i^j} \right)^2 \right] \tag{3}$$

w_i、h_i 分别是预测第 i 个网格的第 j 个锚框的宽高，与之对应的真实值为 $\hat{w}_i^j$、$\hat{h}_i^j$。

L_3 为置信度损失，包括有目标和无目标两种情况，其详细公式如式(4)：

$$\begin{aligned} L_3 = &-\sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{ij}^{obj} \left[\hat{c}_i^j \log\left(c_i^j\right) + \left(1-\hat{c}_i^j\right) \log\left(1-c_i^j\right) \right] \\ &- \lambda_{noobj} \sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{ij}^{noobj} \left[\hat{c}_i^j \log\left(c_i^j\right) + \left(1-\hat{c}_i^j\right) \log\left(1-c_i^j\right) \right] \end{aligned} \tag{4}$$

L_4 为类别损失，其详细公式如式(5)：

$$L_4 = -\sum_{i=0}^{S^2} I_{ij}^{obj} \sum_{c \in classes} \left[\hat{p}_i^j \log\left(p_i^j\right) + \left(1-\hat{p}_i^j\right) \log\left(1-p_i^j\right) \right] \tag{5}$$

3. 基于 F-YOLO V3 模型的人数统计方法

本节首先自建并丰富数据集，提高训练数据的多样性，并通过 K-means 算法重新聚类锚框；然后进行 YOLO V3 特征提取网络以及多尺寸检测算法的改进，提出了 F-YOLO V3 模型；为了衡量检测框定位准确度，增加一个 ADIOU Loss 分支。

3.1. 自建室内人群检测数据集

由于多数情况检测对象人头目标较小，分辨率和信息有限，使得基于深度学习的目标检测算法在常规目标检测数据集上的检测效果并不理想，需要专门针对该检测对象特征的数据库，完成训练和检测任

务。因此，本文首先利用动态数据集和静态数据集相结合自建室内人群检测数据集，并且利用对比度增强、亮度增强等方法进行了数据增强，丰富了数据集，提高了训练数据的多样性。

静态数据集是利用 Python 网络爬虫技术在必应、百度上面爬取室内场景人群的图片，部分图片如图 2 所示。动态数据集采集于北京建筑大学不同教室不同时间的监控视频，视频大小共 20G，格式为 MKV，本实验将视频以 8 秒一帧的标准输出成相应的图像序列，部分图片如图 3 所示。

为丰富数据集，对静态数据集和动态数据集所获得的图片进行数据增强，如对比度增强、亮度增强等，并经过人工筛选得到符合要求的图片组成数据集，最终数据集含有 1000 张图像，只有 1 个类别 person，然后采用开源的 Labelimage 软件对采集到的 1000 张图像中的人头进行标注，得到 1000 个对应的 xml 文件，作为室内人群检测的数据集标签。

Figure 2. Several images on the static dataset
图 2. 静态数据集的部分图片

Figure 3. Several images on the dynamic dataset
图 3. 动态数据集的部分图片

3.2. 聚类候选锚框

锚框是从训练集真实框(ground truth)中统计或聚类得到的几个不同尺寸的框。避免模型在训练时的盲目性，有助于模型快速收敛。设每个网格对应 k 个 anchor，也就是模型在训练时，只会在每一个网格附近找出这 k 种形状。anchor 是对预测的对象范围进行约束，并加入了尺寸先验经验，从而实现多尺度学习目的。

YOLO V3 网络在常规目标数据集 COCO 上通过边框聚类预设了 9 个共 3 类锚框，预设的锚框检测的目标尺寸差距很大，对于普通常规数据集有较好的适应性。但对于一些小目标数据集或者大小尺寸较平均的数据集来说，继续使用这个预设尺寸不利于目标框的收敛，严重影响检测性能。

本实验利用 K-means 算法对自建人群检测数据集所有样本真实框(ground truth)的宽高进行聚类，得到先验框大小，关于锚框数量，原网络输出 3 个尺寸的特征图，所以取了 9 个锚框，由于改进的算法多

输出了 104 × 104 尺寸的特征图，当 k 取 12 时，在自建人群检测数据集中聚类得到这 12 个先验框的尺寸分别是：(11 × 20)，(19 × 32)，(25 × 52)，(37 × 89)，(43 × 56)，(58 × 106)，(68 × 106)，(89 × 125)，(108 × 187)，(110 × 295)，(166 × 230)，(213 × 339)。最后为了提高网络对小目标的召回率和检测的准确度,取消了 13 × 13 尺寸特征图的输出，因此我们将 13 × 13 尺寸特征图所对应的 3 个先验框去掉，最终的 9 个先验框尺寸分别是：(11 × 20)，(19 × 32)，(25 × 52)，(37 × 89)，(43 × 56)，(58 × 106)，(68 × 106)，(89 × 125)，(108 × 187)。

3.3. 改进特征提取网络与多尺度预测的 F-YOLO V3 模型

为了使网络能够获取更多特征信息，增强对模糊或者较小目标检测能力，本文改进了 YOLO V3 算法的特征提取网络与多尺度检测网络，使其充分学习浅层特征，改进的F-YOLO V3网络结构如图 4 所示。

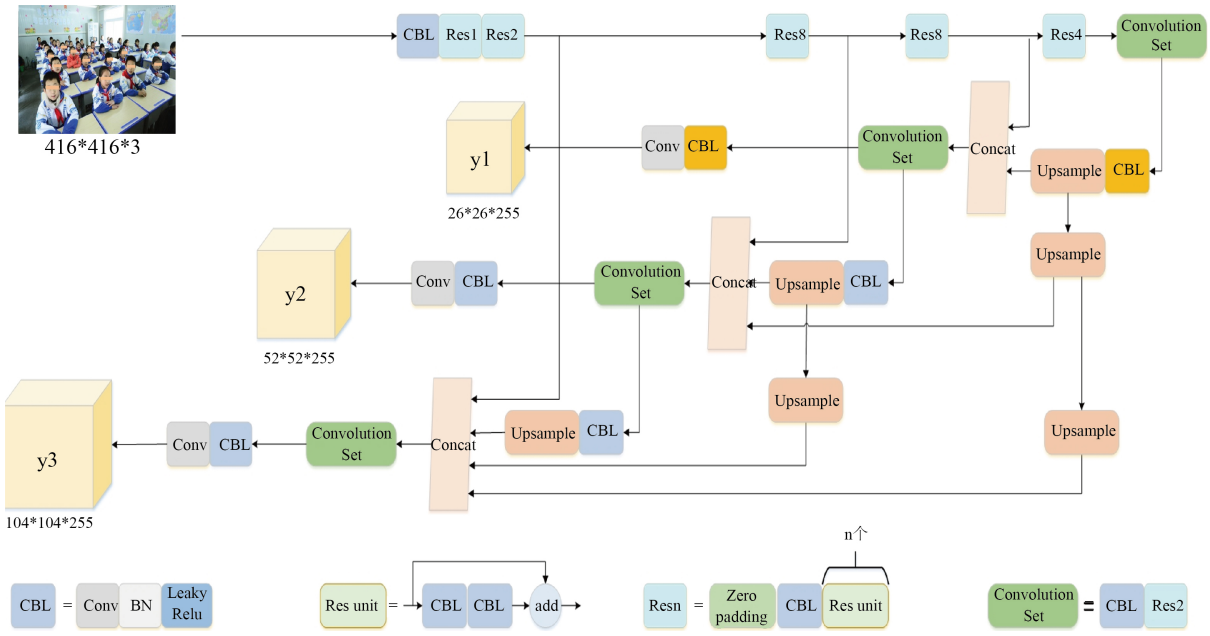


Figure 4. F-YOLO V3 network structure
图 4. F-YOLO V3 网络结构

首先利用低层特征图包含更多特征细节的特点，将原网络输出的 8 倍降采样 52 × 52 的特征图进行上采样，再将得到的结果与 Darknet-53 中第 2 个残差块输出的 4 倍降采样特征图进行拼接，得到 104 × 104 尺寸特征图的输出，可以提取更多的特征信息，提高模糊或者较小目标的检测精度。同时为了提高网络对小目标的召回率和检测准确度，取消了 13 × 13 尺寸特征图输出，最终输出 26 × 26、52 × 52、104 × 104 三种尺寸的特征图。

然后将原网络上采样后的 26 × 26 尺寸特征图再进行 2 倍上采样和 4 倍上采样，得到 52 × 52 和 104 × 104 的特征图与原网络的 52 × 52 和 104 × 104 的特征图进行拼接，同样将原网络上采样后的 52 × 52 尺寸特征图再进行 2 倍上采样，得到 104 × 104 的特征图与原网络的 104 × 104 的特征图进行拼接，这样可以从小目标中得到更细粒度的特征以及位置信息，从而增加目标识别与其位置的准确率。

最后为了增强特征的充分提取，将输出层前的 5 个卷积变成了 1 个卷积和 2 个残差单元，以此减少人头小目标在复杂场景的漏检率，提高网络对小目标的检测率。

3.4. ADIOU Loss

在 YOLO V3 中，smooth L1 Loss 来对检测框的位置做回归，当使用这个 Loss 来做损失函数时，是通过 4 个点回归坐标框的方式，独立的求出 4 个点的 Loss，然后进行相加得到最终的 Bounding Box Loss，这种做法的假设是 4 个点是相互独立的，实际是有一定相关性的，实际评价框检测的指标是使用 IOU，这两者是不等价的，多个检测框可能有相同大小的 smooth L1 Loss，但 IOU 可能差异很大，为了解决这个问题就引入了 IOU Loss。

但 IOU 只是面积比值，当两个框不存在交集时，IOU 为 0，这时网络无法判断两个框之间距离远近，并且从面积比值中无法知道两框的重叠状态，因此本文中增加了一个 ADIOU Loss 分支对预测框和真实框的位置 x、y、w、h 进行计算，ADIOU 增加了一个基于真实框未重叠部分面积的惩罚项；并且针对 IOU 相同但两框相交情况多样的问题，增加了一个基于中心点距离的惩罚项，通过最小化两个框中心的距离来使预测框定位更加准确。

两框之间的重叠度 IOU 定义为：

$$\mathrm{IOU}(B,G)\triangleq\frac{|B\cap G|}{|B\cup G|},\mathrm{IOU}(B,G)\in[0,1] \tag{6}$$

IOU Loss 定义为：

$$\text{IOU Loss}=1-\text{IOU} \tag{7}$$

而 ADIOU 的定义为：

$$\mathrm{ADIOU}(B,G)\triangleq IOU-\frac{|B-B\cap G|}{|C|}-\frac{\rho^2\left(b,b^{gt}\right)}{d^2+1} \tag{8}$$

L_5 为 ADIOU Loss，其详细公式如式(9)：

$$L_5=1-\mathrm{IOU}+\frac{|B-B\cap G|}{|C|}+\frac{\rho^2\left(b,b^{gt}\right)}{d^2+1} \tag{9}$$

其中 B 表示预测框的位置信息，G 表示真实框的位置信息，C 表示包含 B 和 G 的最小矩形框，b 为预测框 B 的中心点，b^{gt} 为真实框 G 的中心点，ρ 为欧氏距离，d 为预测框与真实框重叠部分对角线长度。

因此，F-YOLO V3 模型的总损失函数为：

$$\text{Loss}=L_1+L_2+L_3+L_4+L_5 \tag{10}$$

4. 实验设计

本节将进行三组实验来验证改进的效果，1)将所提的 F-YOLO V3 模型与传统 YOLO V3 模型、鞠[25]改进的 YOLO V3 模型以及 YOLO V5 模型在自建人群检测数据集上进行对比实验；2)消融实验，证明每个改进部分有利于模型的提升；3)将所提的 F-YOLO V3 模型与传统 YOLO V3 模型、成玉荣[21]等人对 Tiny-YOLO V3 改进得到的模型在 SCUT-HEAD 数据集的测试集上进行检测性能的对比。

4.1. 评价指标

本节主要采用主流的目标检测模型的评价指标精确率 Precision、目标召回率 Recall 和平均精度 AP。精确率 Precision 一般指模型检测出来的目标有多大比例是真正的目标物体，定义如式(11)所示。

$$\text{Precision}=\frac{\text{TP}}{\text{TP}+\text{FP}} \tag{11}$$

召回率 Recall 指所有真实的目标有多大比例被我们的模型检测出来了，定义如式(12)所示。

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{12}$$

本文以室内人群检测为例，TP 为正确检测出来的人数，FP 为被错误检出的人数，FN 为没有被检测出来的人数。

以 Recall 为横坐标，Precision 为纵坐标，绘制 P-R 曲线，AP 就是对 PR 曲线上的 Precision 值求均值，对于 PR 曲线可用积分求 AP 的值，定义如式(13)所示，AP@0.5 指将 IOU 阈值设为 0.5；mAP 是 AP 的平均值，c 表示类别数，定义如式(14)所示，mAP 用来衡量模型好坏，值越大说明模型识别的越好，本文只有一个类别 person，因此 AP 就等同于 mAP。

$$\text{AP} = \int_0^1 P(R)\mathrm{d}(R) \tag{13}$$

$$\text{mAP} = \frac{\int_0^1 P(R)\mathrm{d}(R)}{c} \tag{14}$$

$$\text{虚警率} = 1 - \text{Precision} \tag{15}$$

$$\text{漏警率} = 1 - \text{Recall} \tag{16}$$

4.2. 数据集

本文使用了自建室内人群检测数据集以及 SCUT-HEAD 和 Brainwash 两个公开的数据集，下面我们从数据集大小和数据特点等方面分别介绍这 3 个数据集。

自建室内人群检测数据集来源于视频提取和网络爬虫两部分，最终数据集含有 1000 张图像，然后采用开源的 Labelimage 软件对采集到的 1000 张图像中的人头进行标注，得到了 VOC 格式的 xml 文件，作为室内人群检测数据集标签，以 9:1 的比例将数据集标签分为训练样本和测试样本。

SCUT-HEAD 是一个大规模的头部检测数据集，数据集由两部分组成。Part A 包括 2000 个图像，采样自大学教室的监控视频，67321 个注释。Part B 包括 2405 张从互联网上抓取的图片和 43930 个头部注释。我们用 xmin，ymin，xmax 和 ymax 坐标标记每个可见的头部，并确保注释覆盖整个头部，包括被阻塞的部分，但没有额外的背景。我们将 Part A 按照 9:1 的比例分为训练和测试两部分。SCUT-HEAD 数据集遵循 Pascal VOC 标准。

Brainwash 数据集是一个密集人头检测数据集，拍摄的是在一个咖啡馆里出现的人群，包含三个部分，训练集约有 6500 张图像，测试约有 900 张图像。

4.3. 实验环境和训练参数

实验条件：所有实验是基于 Pytorch 和 Python 实现的，深度学习框架为 Darknet 53，配置 Intel Core i7 处理器，内存为 32GB，GPU 为 NVIDIA Quadro M1200，4G 显存，16G 内存。

训练参数设置为：1) 训练迭代次数 epoch 设置为 100；2) 每次迭代训练的图像数目 batch_size 设置为 2；3) 将 batch_size 进行分组后送入网络的 subdivision 设置为 1；4) 网络输入尺寸为 416 × 416；5) 学习率为 0.001；6) 降低参数率 ratio 为 0.25。

5. 实验与结果

5.1. F-YOLO V3 模型性能测试及对比实验

在自建人群检测数据集上，硬件平台为 NVIDIA Quadro M1200，将提出的 F-YOLO V3 模型与传统

YOLO V3 模型、YOLO V5 模型以及复现鞠[25]等人提出的改进的 YOLO V3 模型进行对比实验，实验结果如下表 1 所示。

Table 1. Performance comparison of different algorithms
表 1. 不同算法性能对比

算法	*P*	*R*	AP@0.5	AP@0.5:0.95
YOLO V3	46.53%	69.95%	40.30%	13.12%
YOLO V5	67.20%	72.30%	54.60%	17.30%
Improved-YOLO V3 [25]	44.35%	82.22%	59.42%	21.17%
F-YOLO V3	36.89%	86.19%	63.00%	22.56%

结合表 1 对比不同算法在自建人群检测数据集上的检测精度可以看出 F-YOLO V3 模型相较于传统 YOLO V3 模型在 AP 指标上提高了约 22.7%，虽然 P 值有所降低，但是 R 值有很大的提升，因此漏警率降低了许多。从下面识别检测效果图 5 中就能看出漏检显著减少；相较于 YOLO V5 模型提高了约 13.3%，相较于我们复现的鞠[25]等人提出的改进的 YOLO V3 模型提高了 3.58%，为了进一步体现 F-YOLO V3 模型在精度上的提升，绘制精度趋势图如图 5。

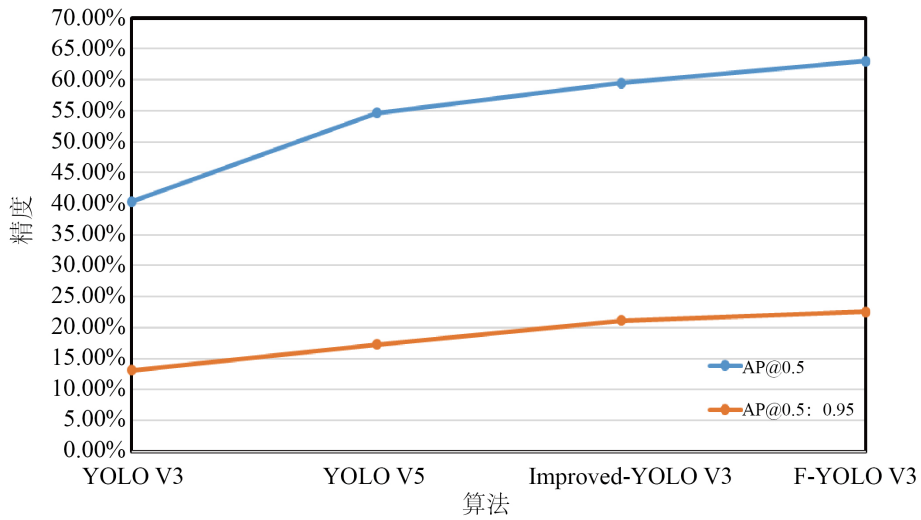


Figure 5. Accuracy trend chart
图 5. 精度趋势图

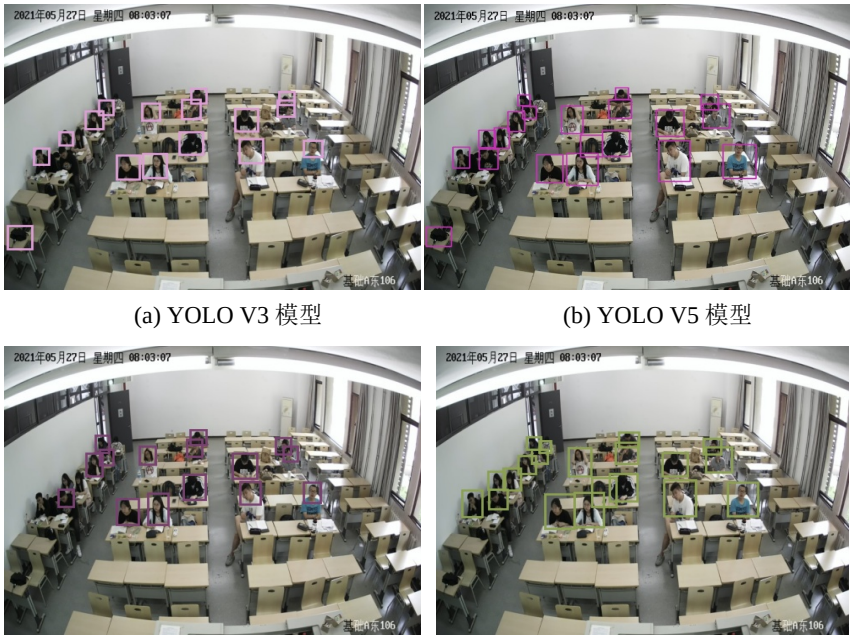
从识别测试效果图 6 中可以看出，YOLO V3 模型和 YOLO V5 模型的都存在一些误检和漏检问题，比如将包检测成人的情况，我们复现的鞠[25]等人提出改进的 YOLO V3 模型存在严重的漏检问题，而 F-YOLO V3 模型可以检测出其他三种算法没有检测出来的目标，并且框的定位准确度提高了很多，对于误检漏检的情况也有显著改善。

5.2. F-YOLO V3 模型的效能评估和分析

本实验将改进的地方分为 A、B、C 三个部分，Part A 为增加一个 ADIOU (Area Distance) loss 分支；Part B 为将原网络输出的 8 倍降采样 52 × 52 的特征图继续进行上采样，与更低层的 104 × 104 尺寸特征

图拼接，增加 104 × 104 尺寸的特征图输出，同时取消 13 × 13 尺寸特征图的输出，并且将输出层前的 5 个卷积变成了 1 个卷积和 2 个残差单元；Part C 为将原网络上采样后的 26 × 26 尺寸特征图再进行 2 倍上采样和 4 倍上采样，得到 52 × 52 和 104 × 104 的特征图与原网络的 52 × 52 和 104 × 104 的特征图进行拼接，同样将原网络上采样后的 52 × 52 尺寸特征图再进行 2 倍上采样，得到 104 × 104 的特征图与原网络的 104 × 104 的特征图进行拼接。

将这三部分在自建人脸检测数据集的训练集上进行训练，并在测试集上对模型进行评估，对比原 YOLO V3 的模型性能，来证明改进的每一部分对于模型的提升效果，实验结果如图 6 和表 2 所示。



(a) YOLO V3 模型　(b) YOLO V5 模型

(c) [25]中的 YOLO V3 模型　(d) F-YOLO V3 模型

Figure 6. Recognition test results
图 6. 识别测试效果图

Table 2. Detection performance under different thresholds
表 2. 不同阈值下的检测性能

算法	*P*	*R*	AP@0.5	AP@0.55	AP@0.6	AP@0.65
YOLO V3	46.53%	69.95%	40.30%	31.86%	25.30%	17.55%
YOLO V3 + A	41.10%	78.07%	46.86%	38.34%	28.11%	19.33%
YOLO V3 + A + B	46.27%	82.71%	62.57%	52.82%	44.15%	31.37%
YOLO V3 + A + B + C (F-YOLO V3)	36.89%	86.19%	63.00%	54.70%	44.72%	31.99%

图 7 显示了四种模型在训练第 50 个和第 100 个 epoch 时训练模型在测试集上返回的 mAP 值，其中测试集上的 IOU 阈值设为 0.5，可以明显看出每一部分的加入都使精度有了一定的提升。表 2 是训练完 100 个 epochs 后不同阈值下的测试结果，P 值虽然有所降低，但是 R 值可以看出是在持续增长的，证明我们想要减少漏警率的目的达到了，图 8 是将表 2 数据绘制成了簇状条形图，最终的 AP 值可以明显看出改进的每一部分都对模型有一定的提升效果。

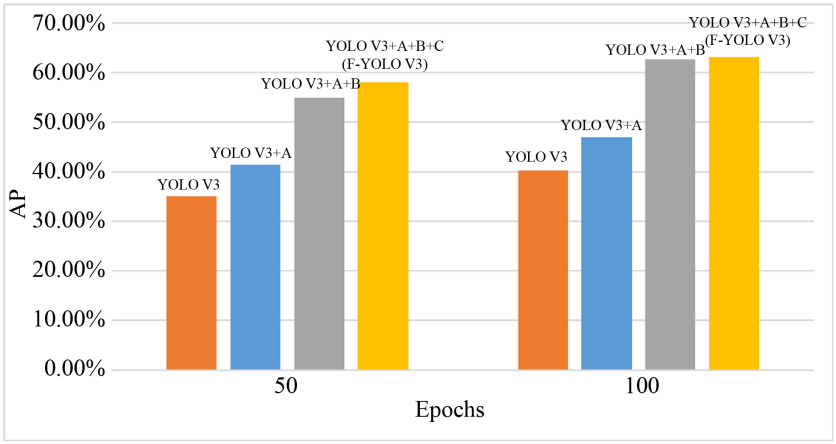


Figure 7. When the four models train the 50th and 100th epoch, they train the map value returned by the model on the test set
图 7. 四种模型在训练第 50 个和第 100 个 epoch 时训练模型在测试集上返回的 map 值

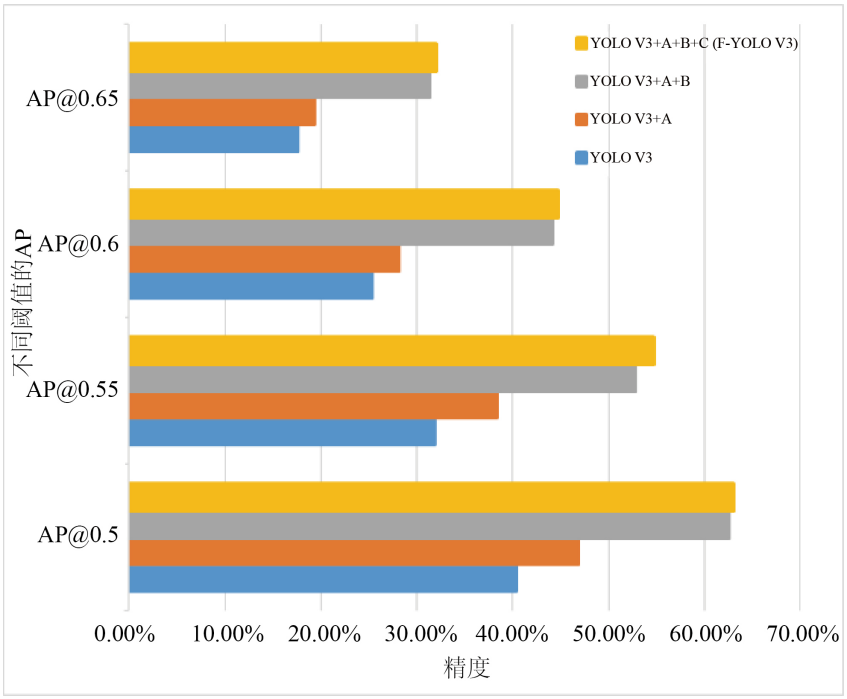


Figure 8. Detection performance under different thresholds
图 8. 模型在不同阈值下的精度

5.3. 在公开数据集上 F-YOLO V3 模型的性能评估

将 F-YOLO V3 模型与传统 YOLO V3 模型、成玉荣[21]等人对 Tiny-YOLOv3 改进得到的模型在 SCUT-HEAD 公开人脸数据集的测试集上进行对比，结果列于表 3 中。

从表 3 可知，F-YOLO V3 模型比传统 YOLO V3 模型平均精度提升了 19.8%，相较于成玉荣[21]等人对 Tiny-YOLOv3 改进得到的模型平均精度提升了 1.27%。

在 Brainwash 数据集上，硬件平台为 NVIDIA GeForce RTX 2080Ti，将 F-YOLO V3 模型与传统 YOLO

V3 模型、我们复现的鞠[25]等人提出的改进的 YOLO V3 模型在 Brainwash 数据集的测试集上进行对比，结果列于表 4 中。

Table 3. Experimental results on SCUT-HEAD public face dataset
表 3. SCUT-HEAD 公开人脸数据集上的实验结果

算法	AP
YOLO V3	61.47%
Improved-Tiny-YOLOv3 [21]	80.00%
F-YOLO V3	81.27%

Table 4. Experimental results on Brainwash dataset
表 4. Brainwash 数据集上的实验结果

算法	AP@0.6
YOLO V3	50.28%
Improved-YOLO V3 [25]	55.83%
F-YOLO V3	58.15%

从表 4 可知，F-YOLO V3 模型比传统 YOLO V3 模型平均精度提升了 7.87%，相较于我们复现的鞠[25]等人提出的改进的 YOLO V3 模型平均精度提升了 2.32%。

从以上两个实验结果可以得出 F-YOLO V3 模型在公共数据集的检测效果也令人满意，进一步说明 F-YOLO V3 模型的有效性。

6. 结论

本文基于改进的 YOLO V3 算法来进行人数统计。首先，利用动态数据集和静态数据集相结合自建数据集，并对部分数据集进行了数据增强；通过 K-means 算法重新聚类锚框；然后进行了 YOLO V3 特征提取网络以及多尺寸检测算法的改进，提出了 F-YOLO V3 模型，以提取更多特征信息，增强对模糊或者较小目标检测能力；最后为了衡量检测框定位准确度，增加一个 ADIOU Loss 分支，最后根据识别出来的人进行统计，输出画面中的实时人数。实验证明，改进后的 YOLO V3 算法在进行室内人数统计时，漏检率和误检率都大大降低，检测准确率有明显提升。

但是，改进后的模型在训练速度、网络结构的复杂度方面还有一定的缺陷，下一步将针对这些缺陷做进一步研究。

基金项目

北京建筑大学科学研究基金(KYJJ2017017, Y19-19, Y18-11)；住房和城乡建设部科学技术计划北京建筑大学北京未来城市设计高精尖创新中心开放课题(No. UDC2019033324, UDC201703332)；北京市教育委员会科学研究计划项目资助(KM202110016001, KM202210016002)。

参考文献

[1] LeCun, Y., Bengio, Y. and Hinton, G. (2015) Deep Learning. *Nature*, **521**, 436-444. https://doi.org/10.1038/nature14539

[2] Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012) ImageNet Classification with Deep Convolutional Neural Net-

work. *Proceedings of the* 25*th International Conference on Neural Information Processing Systems*, Lake Tahoe, 3-8 December 2012, 1097-1105.

[3] Girshick, R., Donahue, J., Darrell, T., *et al*. (2014) Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. *Proceedings of* 2014 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Columbus, 23-28 June 2014, 580-587. https://doi.org/10.1109/CVPR.2014.81

[4] Li, L.-H., Lun, Z.-M., Lian, J., *et al*. (2017) Convolution Neural Network-Based Vehicle Detection Method. *Journal of Jilin University* (*Engineering and Technology Edition*), **47**, 384-391.

[5] Xiong, C.-Z., Shan, Y.-M. and Guo, F.-H. (2017) Image Retrieval Method Based on Image Principal Part Detection. *Optics and Precision Engineering*, **25**, 792-798. https://doi.org/10.3788/OPE.20172503.0792

[6] Li, Y., Qi, H., Dai, J., *et al*. (2017) Fully Convolutional Instane-Aware Semantic Segmentation. *Proceedings of* 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 4438-4446. https://doi.org/10.1109/CVPR.2017.472

[7] Kisantal, M., Wojna, Z., Murawski, J., *et al*. (2019) Augmentation for Small Object Detection. *Proceedings of* 2019 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 1-15. https://doi.org/10.5121/csit.2019.91713

[8] 李航, 朱明. 基于深度卷积神经网络的小目标检测算法[J]. 计算机工程与科学, 2020, 42(4): 649-657.

[9] 易诗, 李欣荣, 吴志娟, 等. 基于红外热成像与改进 YOLO V3 的夜间野兔检测方法[J]. 农业工程学报, 2019, 35(19): 223-229.

[10] Girshick, R., Donahue, J., Darrell, T. and Malik, J. (2014) Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. 2014 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Columbus, 23-28 June 2014, 580-587. https://doi.org/10.1109/CVPR.2014.81

[11] Girshick, R. (2015) Fast R-CNN. 2015 *IEEE International Conference on Computer Vision* (*ICCV*), Santiago, 7-13 December 2015, 1440-1448. https://doi.org/10.1109/ICCV.2015.169

[12] He, K.M., Zhang, X.Y. and Ren, S.Q. (2014) Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, **37**, 1904-1916. https://doi.org/10.1109/TPAMI.2015.2389824

[13] Law, H. and Deng, J. (2018) CornerNet, Detecting Objects as Paired Key-Points. *Proceedings of the European Conference on Computer Vision* (*ECCV*), Munich, 8-14 September 2018, 734-750. https://doi.org/10.1007/978-3-030-01264-9_45

[14] Ren, S.Q., He, K.M., Girshick, R., *et al*. (2017) Faster R-CNN, towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **39**, 1137-1149. https://doi.org/10.1109/TPAMI.2016.2577031

[15] He, K.M., Gkioxari, G., Dollar, P., *et al*. (2017) Mask R-CNN. 2017 *IEEE International Conference on Computer Vision* (*ICCV*), Venice, 22-29 October 2017, 2980-2988. https://doi.org/10.1109/ICCV.2017.322

[16] Redmon, J., Divvala, S., Girshick, R., *et al*. (2016) You Only Look Once, Unified, Real-Time Object Detection. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 779-788. https://doi.org/10.1109/CVPR.2016.91

[17] Redmon, J. and Farhadi, A. (2017) YOLO9000, Better, Faster, Stronger. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 6517-6525. https://doi.org/10.1109/CVPR.2017.690

[18] Redmon, J. and Farhadi, A. (2018) YOLOv3, an Incremental Improvement. https://arxiv.org/abs/1804.02767

[19] Liu, W., Anguelov, D., Erhan, D., *et al*. (2016) SSD, Single Shot Multibox Detector. In: Leibe, B., Matas, J., Sebe, N., *et al*., Eds., *Lecture Notes in Computer Science*, Springer, Cham, 21-37. https://doi.org/10.1007/978-3-319-46448-0_2

[20] 陈晓. 基于目标检测的视频人数统计算法研究[D]: [硕士学位论文]. 哈尔滨: 哈尔滨工业大学, 2019.

[21] 成玉荣, 胡海洋. 基于改进 Tiny-YOLOv3 的人数统计方法[J]. 科技创新导报, 2020, 17(10): 4-5+8.

[22] 郑国书, 朱秋煜, 王辉. 基于深度学习 SSD 模型的视频室内人数统计[J]. 工业控制计算机, 2017, 30(11): 48-50.

[23] He, K.M., Zhang, X.Y., Ren, S.Q., *et al*. (2016) Deep Residual Learning for Image Recognition. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 770-778. https://doi.org/10.1109/CVPR.2016.90

[24] Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012) ImageNet Classification with Deep Convolutional Neural Networks. *Proceedings of the* 25*th International Conference on Neural Information Processing Systems*, Lake Tahoe, 3-8 December 2012, 1097-1105.

[25] 鞠默然, 罗海波, 等. 改进的 YOLO V3 算法及其在小目标检测中的应用[J]. 光学学报, 2019, 39(7): 0715004.