计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240400192

一种基于目标检测的轨道交通上下客区客流指引方法

乐凌志[1,2]　翟江涛[2]　俞　铭[1]　孙同庆[2]
1 南京国电南自轨道交通工程有限公司　南京 210000
2 南京信息工程大学电子与信息工程学院　南京 210044

摘　要　针对等待期屏蔽门前乘客占用下客区的情况,提出了一种基于目标检测的上下客区客流指引方法。首先针对屏蔽门前场景中乘客的形状特征对目标检测网络进行改进,提出 MCA-YOLOv5s 网络模型。然后通过智能门楣系统的安装高度和真实场景中上下客区的范围计算出摄像头的视场角大小和安装角度,确保拍摄的图像能够准确划分出上下客区。最后分别对上下客区中的乘客进行密度估计并设计对应密度值的客流分配策略,通过智能门楣终端上的扬声器进行指引。通过在真实场景中进行测试,验证了所提方法能够快速准确地估计乘客密度。

关键词:目标检测;乘客密度;客流分配
中图分类号　TP391

Object Detection-based Method for Guiding Passenger Flow in Boarding and Deparking Areas of Rail Transit

LE Lingzhi[1,2],ZHAI Jiangtao[2],YU Ming[1] and SUN Tongqing[2]
1 Nanjing Guodian Nanzi Rail Transit Engineering Co.,Ltd.,Nanjing 210000,China
2 School of Electronics and Information Engineering,University of Information Science & Technology,Nanjing 210044,China

Abstract　To address the situation where passengers occupy the alighting area while waiting in front of the platform screen doors,this paper proposes a passenger flow guidance method based on object detection. Firstly,an improved MCA-YOLOv5s network model is proposed by enhancing the shape features of passengers in front of the platform screen doors for object detection. Then the field of view angle and installation angle of the camera are calculated based on the mounting height of the intelligent door lintel system and the range of the alighting area in real scenes to ensure accurate division of the alighting and boarding areas in captured images. Subsequently,passenger density estimation is conducted for the alighting and boarding areas,and corresponding passenger flow distribution strategies are designed based on the estimated density values,with guidance provided through speakers on the intelligent door lintel terminal. Through testing in real scenarios,the effectiveness of this method in rapidly and accurately estimating passenger density is validated.

Keywords　Object detection,Passenger density,Passenger flow distribution

1　引言

针对屏蔽门前上下客区乘客排队的乱象,传统的处理方法以人工指引和张贴静态客流指引标识为主,但是站台范围太广,传统的人工客流指引方法无法对每个乘客进行准确的定位和跟踪。随着人工智能技术在轨道交通领域的应用,目前主流的客流指引方法是先估计出每个屏蔽门前乘客的密度,然后再进行指引的方法。目前国内外关于轨道交通的客流密度估计技术通常包括以下几种方式:无线电波检测、人工观察统计和基于视频分析的检测方法等。基于视频分析的客流密度估计通常借助摄像头实时监控目标区域,结合计算机视觉、图像处理等技术对获取的视频进行分析,从而计算出客流密度。针对轨道交通中的乘客进行密度估计一般选择人体的局部作为检测的目标,本文的摄像头安装在屏蔽门门楣处,对真实场景的屏蔽门前乘客数据进行分析后,选择遮挡较小的头部作为检测目标。

当前的目标检测算法主要包括传统检测方法和基于深度学习的方法,传统的方法主要采用手工设计提取局部特征或全局特征,例如 SIFT(Scale-Invariant Feature Transform)、HOG(Histogram of Oriented Gradients)、LBP(Local Binary Patterns)以及利用 Haar 特征的方法等,但这些方法只能适应简单场景,需要消耗大量的人力进行时间选择和特征设计,在复杂场景中效果较差。随着卷积神经网络(Convolutional Neural Network,CNN)在图像处理中的广泛应用,计算机视觉检测技术得到了极大的发展,开始从大量的数据中学习到目标特征后再进行分类,国内外基于 CNN 相继提出了 AlexNet,VGG(Visual Geometry Group),GoogLeNet 系列和 ResNet(Residual Network)等网络模型,能够提取到更加丰富的

基金项目:国家自然科学基金(U21B2003);江苏省产业前瞻与关键核心技术项目(BE2022075)
This work was supported by the National Natural Science Foundation of China(U21B2003) and Industry Outlook and Key Core Technology Projects in Jiangsu Province(BE2022075).
通信作者:乐凌志(lingzhi-le@sac-china.com)

目标特征。根据神经网络设计的目标检测类型可以分为two-stage检测算法和one-stage检测算法两类。two-stage算法将图像候选区域和卷积神经网络进行融合，使用神经网络提前在输入图像的生成区域中创造一个目标分类器，然后进行分类和特征提取。2014年，Ross Girshick等提出了R-CNN检测算法，采用选择性搜索(Selective Search，SS)算法来生成候选框作为输入，极大地提高了目标检测算法的检测性能，但是检测速度受限。针对R-CNN检测算法中的不足，作者后续推出了Fast R-CNN[1]、Faster R-CNN[2]算法，但是仍需要提前获取大量的候选框，速度依旧是瓶颈，实用性较低。在Faster R-CNN的基础上，He等[3]提出了Mask R-CNN算法，使用一个额外的分支预测目标的掩膜与检测到的物体进行融合。One-stage算法是基于回归的目标检测方法，不需要生成候选框，直接对初始的目标进行检测，加快了特征的提取速度。Liu等提出的SSD算法[4]利用多尺度特征图进行目标检测，Redmon等随后提出的YOLO9000和YOLOv3算法采用了更深的骨干网络进行特征提取，Bochkovskiy等提出的YOLOv4[5]算法在YOLOv3的基础上融入当时许多先进的网络结构。Glenn等提出YOLOv5算法，在Neck端使用FPN+PAN结构，训练过程中使用多种有效的策略，极大地提升了目标检测算法的检测速度和精度。最近YOLOvX[6]，YOLOv6[7]和YOLOv7[8]算法也相继提出，算法的性能得到了提升，但是尚未得到成熟运用。

2 基于MCA-YOLOv5s的轻量级目标检测算法

复杂网络模型通常具有较大的参数量，部署到嵌入式设备中将面临占用空间大和检测速度慢的问题，难以满足地铁智能门楣终端大规模监控系统低延迟和快速响应的需求。同时每个屏蔽门智能门楣终端部署需要考虑到工程成本的实际需求。因此，本文选择速度快、精度高以及适合模型部署的单阶段目标检测算法YOLOv5s为基础算法，针对地铁屏蔽门前场景进行改进，提出了一种轻量级目标检测算法MCA-YOLOv5s。该算法具有更快的推理速度，能够有效地节省地铁站内硬件部署资源，更满足地铁屏蔽门前的客流监测和管理的工程实际需求。

2.1 整体算法框架

本文算法在YOLOv5s-6.0版本的基础上进行改进，YOLOv5s网络结构分为输入端、Backbone、Neck和Head。输入端采用自适应图片缩放技术和Mosaic数据增强以及k-means算法处理输入的图像。在Backbone部分中，特征图首先会使用一个较小的卷积核对特征进行初步的提取，接着通过4层C3模块生成不同尺寸的特征图，最后使用空间金字塔池化结构SPPF(Spatial Pyramid Pooling-Fast)融合不同感受野的特征图。Neck部分采用FPN+PAN[9]结合的路径聚合网络架构，加强网络特征的融合能力，其中下采样层使用了C3模块，用于在保持特征图维度的同时将输入特征图的尺寸减小一半。Head检测层分别解码预测3种不同尺寸的特征图，使用NMS(Non-Maximum Suppression)非极大值抑制算法获取目标最优预测框，输出预测框和类别位置信息。

为优化智能门楣终端大规模监控场景下的乘客检测性能，对YOLOv5s网络结构进行改进，提出了一种基于MCA-YOLOv5s的轻量级目标检测算法。其整体框架如图1所示，对于输入的乘客图像，首先采用轻量级模块Mobilenetv3[10]重构YOLOv5s的主干网络，减少模型的体积和参数量，实现网络模型轻量化处理，并用PConv(Partial Convolution)[11]代替深度可分离卷积中的DWConv(Depthwise Separable Convolution)，减少冗余计算和内存访问，提高网络的计算速度。为了增强网络各层的特征融合能力，在特征融合模块的C3层中融入CA(Coordinate Attention)[12]注意力模块，使模型更加关注目标的位置信息，提高对目标位置的定位能力；最后将损失函数CIoU(Complete IoU)替换为Alpha IOU[13]以增加High Loss目标的权重和边界框的回归精度，优化模型整体性能。由于经过相机视场校准后的智能门楣终端安装在距离地面2.5 m高的屏蔽门门楣处，相机的拍摄角度为俯视向下60°，因此拍摄的上下客区内的乘客具有很大的图像尺度，为此在原先输出端结构的基础上去掉一个小尺度的检测层，在保证大尺度图像检测精度的同时降低整体网络的计算复杂度。

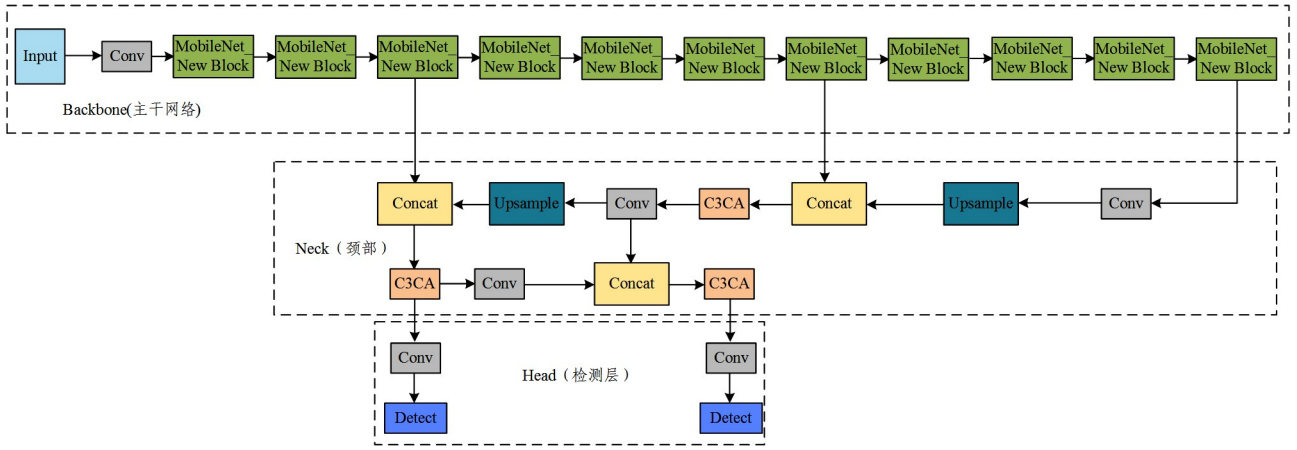


图1 MCA-YOLOv5s检测算法

Fig. 1 MCA-YOLOv5s detection algorithm

2.2 改进的轻量级主干网络结构

本文提出一个新的轻量级主干网络P-Mobilenetv3，由11个MobileNet_New Block组成。其中的MoblieNetv3采用MoblieNetv1[14]和MoblieNetv2[15]中提出的深度可分离卷积和逆残差结构，在此基础上更新Block，加入SE(Squeeze and Excitation)[16]模块，利用h-swish代替swish激活函数，进一

步提高了计算速度和模型性能。MoblieNetv3 网络中的 Block 网络结构如图 2 所示，主要包括通道可分离卷积和 SE 通道注意力机制以及残差网络结构，其核心是使用深度可分离卷积代替传统卷积层，将传统卷积层拆分成逐通道卷积（DWConv）和逐点卷积（PWConv）。

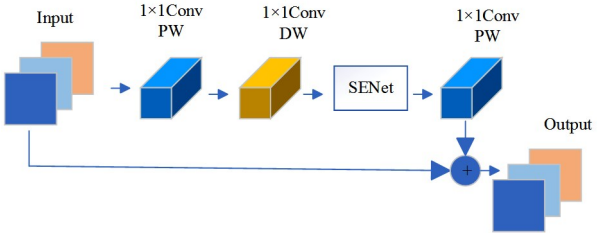


图 2　Block 网络结构

Fig. 2　Block network structure

Mobilenetv3 通过 NAS 搜索全局网络结构，分为 Large 和 Small 两种版本，主要的不同在于经过卷积升维后的通道数量以及网络中的 Block 使用次数，本文采用 Mobilenetv3-Small 模型进行实验。

为了减少逐通道卷积中的冗余计算和内存访问的数量，本章使用 PConv（Partial Convolution）替换 DWConv，以更好地平衡检测延迟（Lantency）和浮点运算（FLOPs）之间的联系，联系计算式如下：

$$Lantency=\frac{FLOPs}{FLOPS} \tag{1}$$

其中，$FLOPS$ 表示每秒浮点运算的缩写，度量有效的计算速度，PConv 可以缓和网络进行 $FLOPs$ 时内存访问频繁造成 $FLOPS$ 减小的副作用，在降低 $FLOPs$ 的同时优化 $FLOPS$，尽可能多地使用设备的计算能力，实现更好的低延迟效果。PConv 的工作原理如图 3 所示。

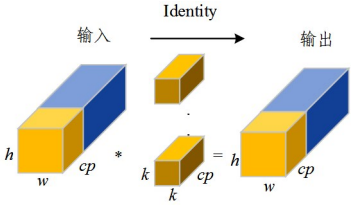


图 3　PConv 工作原理图

Fig. 3　PConv working principle diagram

在 PConv 结构中，只需要使用部分输入图像的通道与标准卷积结合进行特征的提取，其余通道保持不变，如果内存访问是连续或者规则的，使用第一个或最后一个连续的通道作为计算代表与整个特征图进行融合，PConv 的内存访问数量如下：

$$h\times w\times 2\,c_p+k^2\times c_p^2\approx h\times w\times 2\,c_p \tag{2}$$

其中，h 和 w 分别为输入矩阵的宽高，C_p 是常规卷积作用的通道数，k 为卷积核的大小。在实际实现过程中，C_p 一般设置为常规矩阵的 1/4，其余通道数不参与计算，而 DWConv 虽然能降低计算的 $FLOPs$，但是会增大通道数来弥补精度的下降，一般通道数会增大为常规卷积的 6 倍，导致 C_p 会特别大。因此，PConv 相比与 DWConv 能够极大地减小内存访问的数量和计算冗余。

对于其余不参与计算的通道，PConv 层中没有进行简单的删除，而是使用 PWConv 进行剩余通道特征的进一步提取，PWConv 可以提取所有通道特征信息流，充分完整地捕获所有通道的特征信息。PConv 与 PWConv 组合成新的结构 New Block，新主干网络 P-Mobilenetv3 结构如表 1 所列。

表 1　P-Mobilenetv3 网络结构

Table 1　P-Mobilenetv3 network architecture

Input	Operator	exp size	SE	AF	stide
$640^2\times3$	conv2d,3×3	—	—	HS	2
$320^2\times8$	New Block,3×3	16	√	RE	2
$160^2\times8$	New Block,3×3	72	—	RE	2
$80^2\times16$	New Block,3×3	88	—	RE	1
$80^2\times16$	New Block,5×5	96	√	HS	2
$40^2\times24$	New Block,5×5	240	√	HS	1
$40^2\times24$	New Block,5×5	240	√	HS	1
$40^2\times24$	New Block,5×5	120	√	HS	1
$40^2\times24$	New Block,5×5	144	√	HS	1
$40^2\times24$	New Block,5×5	288	√	HS	2
$20^2\times48$	New Block,5×5	576	√	HS	1
$20^2\times48$	New Block,5×5	576	√	HS	1

P-Mobilenetv3 网络对于输入网络的图像，采用 Mosaic 数据增强统一调整为 640×640 的尺寸，operator 表示使用表中的结构依次对输入的图像进行操作，首先采用通道数为 3 的卷积进行处理，然后依次使用 New Block 结构处理图像，该部分结构会降低特征图的尺寸，最终降低为 20×20，exp size 表示对特征图进行升维操作，最终维度升高到 576，整个主干网络中一共包含 11 个 New Block，SE 表示是否使用该结构，HS 为 hard-swish 激活函数，RE 为 ReLU（Rectified Linear Unit）激活函数，stride 表示 operator 的过程中卷积的步长为 2，表示对特征图的尺寸进行压缩。

2.3　针对大尺度人头检测和定位的改进

针对乘客头部在摄像头中的大尺度成像特点以及位置坐标信息在密度估计中的重要性，本文在 yolov5s 特征融合阶段的 C3 模块中融入坐标注意力机制 CA（Coordinate Attention），形成 C3CA 模块，使得模型能够更加关注乘客的位置信息，提高对目标定位的能力，捕获不同通道之间的联系。最后在颈部多尺度目标检测结构中，去掉最后一个小尺度检测层，只使用 40×40 和 80×80 的检测层进行乘客头部的检测，使得模型更加关注地铁站内距离屏蔽门较近的上下客区域，同时还可以加快模型的计算速度，提升网络模型检测乘客的整体性能。

CA 注意力机制相比其他主流的注意力机制，如 SENet，ECA（Efficient Channel Attention）[17]，CBA（Convolutional Block Attention Module）[18]，同时考虑了通道维度和空间维度的信息，并把位置信息嵌入到通道注意力中，有效地解决了空间维度存在的长距离依赖的问题，而且还避免了大量的计算，适合嵌入到轻量化网络中。CA 注意力机制的具体流程如图 4 所示。

CA 注意力机制首先进行信息嵌入操作。针对输入特征图的每一个通道信息，利用尺寸为（H，1）和（1，W）的卷积核对每个通道的水平方向和垂直方向进行全局平均池化操作，聚合两个空间方向的特征，输出具有方向感知的特征图，能够精确地保留位置信息。其中水平方向得到 $H\times1\times C$ 的信息特征图的计算式为：

$$Z_c^h(h)=\frac{1}{W}\sum_{0\leqslant i\leqslant w}x_c(h,i),Z_c^h\in R^{C*H*1} \tag{3}$$

垂直方向得到 $1\times W\times C$ 的信息特征图计算式为：

$$Z_c^w(w)=\frac{1}{H}\sum_{0\leqslant i\leqslant H}x_c(j,w),Z_c^w\in R^{C*1*W} \tag{4}$$

接着进行 CA 注意力生成操作。沿着空间维度对生成的特征图 Z_c^h 和 Z_c^w 进行级联操作，把水平方向和垂直方向的特征级联为全局特征。再使用 1×1 的卷积和激活函数进行 F1 变换。然后在空间维度使用分片操作得到两个单独的注意力张量 g^h 和 g^w，使用两个 1×1 的卷积将张量的通道数变换为和输入相同的通道数。F1 变换、g^h 以及 g^w 分别表示为：

$$f=\delta(F_1([Z^h,Z^w])),f\in R^{\frac{C}{r}*1*(H+W)} \tag{5}$$

$$g^h=\sigma(F_h(f^h)) \tag{6}$$

$$g^w=\sigma(F_w(f^w)) \tag{7}$$

最后使用广播变换把 g^h 和 g^w 拓展到 $C\times H\times W$ 维度，对特征图进行矫正，得到注意力特征。CA 注意力机制输出的最终表达式为：

$$y_c=x_c\times g^h\times g^w \tag{8}$$

CA 注意力机制融入到 C3 模块中只会引入少量的参数，而且不会增加网络的总层数，运用在本文数据集中可以提高整个网络的性能，具体实现如图 5 所示。其中 C3 模块由 3 个 CBS(conv+Batch Normalization+SiLU)模块和 NottleNeck 模块组成，在 C3 模块主通道的 CBS 卷积之后接入 CA 注意力模块组合成 C3CA 模块。其中，CBS 为带有 BN 和激活函数 Silu 的卷积核大小为 1×1 的卷积，BottleNeck 为两个卷积核大小分别为 1×1 和 3×3 的 CBS 卷积融合而成。

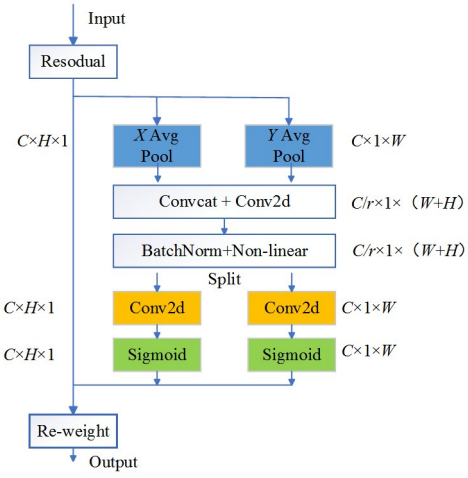


图 4 CA 注意力机制流程

Fig. 4 CA attention mechanism process

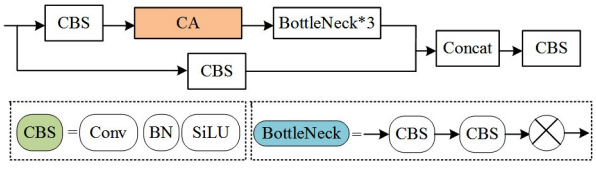


图 5 CA 融合 C3 模块

Fig. 5 C3 fusion to CA module

YOLOv5s 的检测层用来对主干网络提取的特征图进行多尺度目标检测，将提取的特征图经过不同数量的卷积模块进行降维和压缩成不同的特征图尺寸，再将其中不同层级的特征图进行融合，输入到检测层得到不同尺寸的目标检测结果。

由于本文的检测目标是距离智能门楣终端摄像头很近的上下客区域内的乘客头部，该区域内乘客的头部在摄像头成像后在图像中占有很大的尺寸，而距离摄像头较远的地铁站内区域中的乘客并不在统计范围内，因此本文在 YOLOv5s 网络结构的基础上去掉最后一个 20×20 的小尺寸检测头，形成只有 2 个检测头的网络结构，只关心距离较近的大尺寸乘客头部目标。

2.4 损失函数改进

在神经网络的训练过程中，一个好的损失函数应有效地衡量模型预测与实际目标之间的差异，并通过逐步优化权重和参数来最小化预测值与真实值之间的差异，引导模型朝着更好的方向优化，加速模型的收敛和性能提升。

YOLOv5s 的 6.0 版本网络的损失函数 CIOU 是在 GIOU(Generalized Intersection over Union)损失函数的基础上进行的改进。其中 GIOU 损失函数考虑到真实框和预测框不相交的情况下，梯度恒为 0 无法反向传播的问题，提出了使用最小外接矩阵包含真实框和预测框[19]。但是在对行人进行检测时，经常会出现如图 6 所示的真实框与预测框完全重叠在一起的情况，这会导致损失收敛速度变慢。CIOU 损失函数进一步的优化了检测框，同时考虑到预测框与真实框的重叠面积、中心点聚集、长宽比之间的差异[20]，从而使得目标框的发散减少，回归更加稳定。CIOU 的损失函数的定义如下：

$$L_{\text{loss}}=1-IoU+\frac{\rho^2(b,b_{gt})}{c^2}+\gamma u \tag{9}$$

其中，u 是衡量长宽比一致性的相似系数，为主要的改进点，计算式如下：

$$u=\frac{4}{\pi^2}\left(\arctan\frac{w^{gt}}{h^{gt}}-\arctan\frac{w}{h}\right)^2 \tag{10}$$

γ 为调节因子，计算式为：

$$\gamma=\frac{u}{(1-IoU)+u} \tag{11}$$

其中，L_{loss} 为 CIOU 的损失；IOU，ρ，b_{gt} 和 b 以及 c 分别为真实框与预测框的交并比、中心点之间的距离、中心点以及并集部分对角线的长度；w 和 h 分别为预测框的高度与宽度，w^{gt} 和 h^{gt} 分别为真实框的高度和宽。

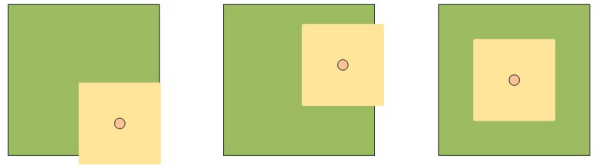


图 6 3 种不同重叠情况

Fig. 6 Three different overlapping situations

本文使用 Alpha IoU 损失函数对 YOLOv5s 中的 CIoU 损失函数进行进一步的优化，增大训练过程中 High IoU 目标的梯度和损失，并解决某些 High IoU 目标回归精度较差的情况，加快收敛速度。其计算式如式(12)所示：

$$L_{\alpha\text{-CIoU}}=1-IoU^{\alpha}+\frac{\rho^{2\alpha}(b,b_{gt})}{c^{2\alpha}}+(\gamma u)^{\alpha} \tag{12}$$

该函数通过在 CIoU Loss 中使用 Box-Cox 变换引入额外的 Power 正则化参数 α 得到。通过调节 α 的数值，式(12)可以概括出包括 GIoU，DIoU(Distance-IoU)和 CIoU 等基于 IoU 的损失，因此能够应用于不同的 bbox 回归精度，对噪声 bbox 有更强的鲁棒性。当 α 大于 1 时，该公式将分配给 High IoU 目标更多的损失权重，有利于模型更好地识别 High IoU 目标，提高 High IoU 边界框的回归精度以及模型整体检测性能。经过实验对比，α 参数取 3 的情况下模型对 High IoU 目标的检测效果最好。

3 上下客区人群密度估计以及客流分配策略设计

本章对目标检测算法检测和定位到的乘客进行密度估计。首先需要根据智能门楣终端的安装高度以及真实场景中地铁屏蔽门前上客区和下客区的范围确定相机的参数和安装的角度，确保乘客被正确地统计在对应的区域中。然后对每个区域的乘客进行密度估计，在使用上节提出的 MCA-YOLOv5s 算法检测到图像中上下客区每个乘客头部的位置信息后，将每一帧图像中检测到的乘客头部像素位置与划分的上下客区域范围进行比对，分别获得上客区和下客区的乘客密度估计值。最后通过设置的针对上下客区不同乘客密度值情况下的管控策略，对屏蔽门前候车乘客进行指引。

3.1 划区域人群密度估计

用于人群密度估计的智能门楣终端安装在距离地面高度 2.5 m 处的屏蔽门门楣处，其中的相机以俯视的角度捕捉屏蔽门前区域的画面。该相机需要能够捕捉到与屏蔽门垂直距离 5 m 远、水平距离 3 m 远处地方的画面。此外，相机的采集最大视场需要包括屏蔽门水平线与垂直距离 5 m 远处最高点之间的夹角，采集的范围大致如图 7 所示。

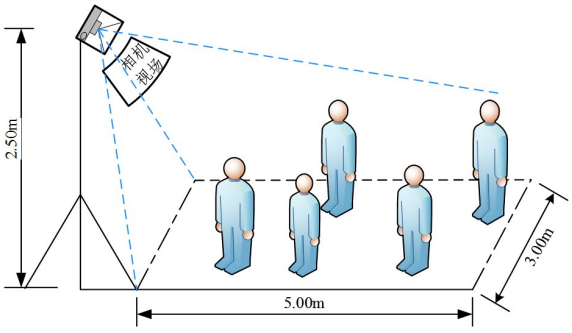


图 7 相机采集范围

Fig. 7 Camera acquisition range

该图为模拟屏蔽门前上下客区场景图像，其中虚线为地铁上客区和下客区的范围。拍摄相机需要满足足够的视场角（*FOV*）范围和最远的采集距离，保证每个上下客区的乘客头部都能被拍摄到。其中 *FOV* 包括对角线 *FOV*，垂直 *FOV*，水平 *FOV*，它指镜头的中心点与成像平面对角线两端所形成的夹角，它决定了镜头所能捕捉到的画面范围，其成像原理如图 8 所示。

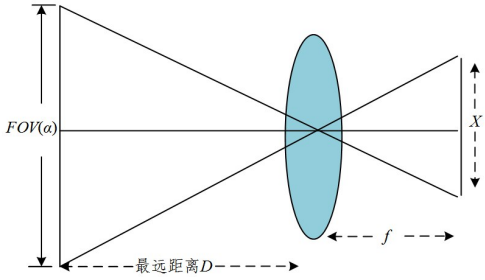


图 8 FOV 成像原理

Fig. 8 FOV imaging principle

其中，x 为相机的传感器（CMOS）尺寸，f 为焦距，α 为视场角，由于每一个摄像头的传感器宽度是固定的，因此视场角的大小和焦距的大小有关，它们之间的对应关系如式（13）所示：

$$FOV=2\arctan\left(\frac{x}{2f}\right) \tag{13}$$

根据式（13）可知，相机的焦距越小，视场角越大。智能门楣终端支持树莓派摄像头，其中的传感器尺寸为 6.35 mm，约为 3.2 mm×2.4 mm，分别表示传感器尺寸的宽度和高度，计算 *FOV* 时，x 表示的是传感器的对角线尺寸，其计算式为：

$$x=\sqrt{\text{宽度}^2+\text{高度}^2} \tag{14}$$

带入参数可得树莓派摄像头传感器的对角线尺寸约为 4 mm，该传感器是一种常见的小型图像传感器尺寸，在一些便携式摄像头、网络摄像头和移动设备中使用。尽管尺寸较小，但这种尺寸的传感器足以采集到智能门楣终端需要的图像质量。通过传感器尺寸 x 和最远工作距离 *WD* 以及最物体的高度 H 能够得到相机的焦距，它们之间的关系如下：

$$f=\frac{x*WD}{H} \tag{15}$$

把传感器尺寸 x 的高度值以及需求的拍摄距离和高度参数值带入可得到焦距值约为 4.8 mm，将该焦距值和传感器的对角线尺寸 x 带入式（12）可得到对角线 *FOV* 的角度约为 45°。由于焦距越小，相机可以更好地覆盖检测范围，因此本文选择型号为 RPI Camera（D）的树莓派摄像头，该相机支持 500 万高清像素的图像处理，感光元件是 OV5647，焦距为 3.4 mm，对角视场角（*FOV*）为 66°，支持 720P、60FPS 视频的采集。该相机拍摄的范围能够满足智能门楣终端的采集需求，拍摄具有高质量和高帧率的图像质量，并且不会拍摄到上下客区范围外的乘客头部，能够提供更准确的乘客密度估计结果。

由于该相机的安装高度固定为 2.5 m，期望相机能够看到距离门前 5 m 的区域，因此相机的安装俯视角度可以用反正切函数得到，最后通过在实际环境中对相机进行测试，对角度进行微调，选择俯视角度 60°作为相机的安装角度。

乘客密度估计的方法主要分为基于目标检测和基于回归模型的方法。相对于基于回归模型的方法，基于目标检测的 MCA-YOLOv5s 算法能够自动学习到乘客头部特征，获取乘客头部的位置信息，不需要人工选取特征，整体具有快速的推理速度，并且不会丢失图片的局部区域特征。又因为智能门楣终端摄像头安装在距离地面 2.5 m 高的门楣处，并且采集角度正对屏蔽门前乘客，距离乘客很近，所以乘客的头部特征会很明显并且尺寸较大。针对上述原因，本文使用目标检测算法先对乘客的头部进行检测，然后根据头部位置信息进行人群密度估计。

本文使用黄色线段将图像按照真实场景比例自动划分为上客区和下客区，在使用 MCA-YOLOv5s 目标检测算法检测到乘客头部位置信息后，分别将每个区域中的乘客头部数量进行统计，分别得出上客区和下客区的乘客密度，划分效果如图 9 所示，其中黄色底线为候车安全线，红框为上下客区域范围。

图 9 上客区和下客区划分示意图（电子版为彩图）

Fig. 9 Diagram of the division of boarding and deparking areas

3.2 基于密度值的上下客区管控

根据统计出的上下客区域不同的密度情况，智能门楣终端通过设置不同的乘客密度阈值，实现科学、有效的客流指引方案。假设上客区密度为 a，下客区密度为 b，总密度为 c，具体的管控流程如图 10 所示。

不同的乘客密度值情况下，本文一共设置了 5 种管控措施。

1）首先对上下客区域的乘客总密度进行判断，针对总密度 c 大于屏蔽门前最大容量阈值的情况，说明此时上客区和下客区都拥挤。通过智能门楣终端扬声器指引乘客尽量前往其他屏蔽门前候车。

2）对于总密度 c 小于屏蔽门前最大容量阈值，上客区密度 a 大于设置的二级阈值的情况，说明此时上客区拥挤。在保障上客区乘客安全的前提下，允许下客区少量的乘客侵占，同时指引乘客前往其他屏蔽门前候车。

3）对于上客区密度 a 小于设置的一级阈值的情况，说明此时上客区不拥挤。若下客区有侵占乘客，指引其前往上客区候车。

4）对于上客区密度 a 在一级阈值和二级阈值之间，同时下客区乘客 b 大于阈值的情况，说明此时上客区较拥挤，下客区拥挤。在保障下客区乘客安全的前提下，允许下客区有少量侵占乘客，同时指引下客区侵占的乘客前往上客区候车或者前往其他屏蔽门前候车。

5）对于上客区密度 a 在一级阈值和二级阈值之间，同时下客区乘客 b 小于阈值的情况，说明此时上客区较拥挤，下客区不拥挤。若下客区有侵占乘客，指引其前往上客区候车。

在划定的上下客区域内，通过实际测试，最大容量阈值一般设置为 20 人，上客区二级阈值为 12，一级阈值为 6，下客区阈值为 4。

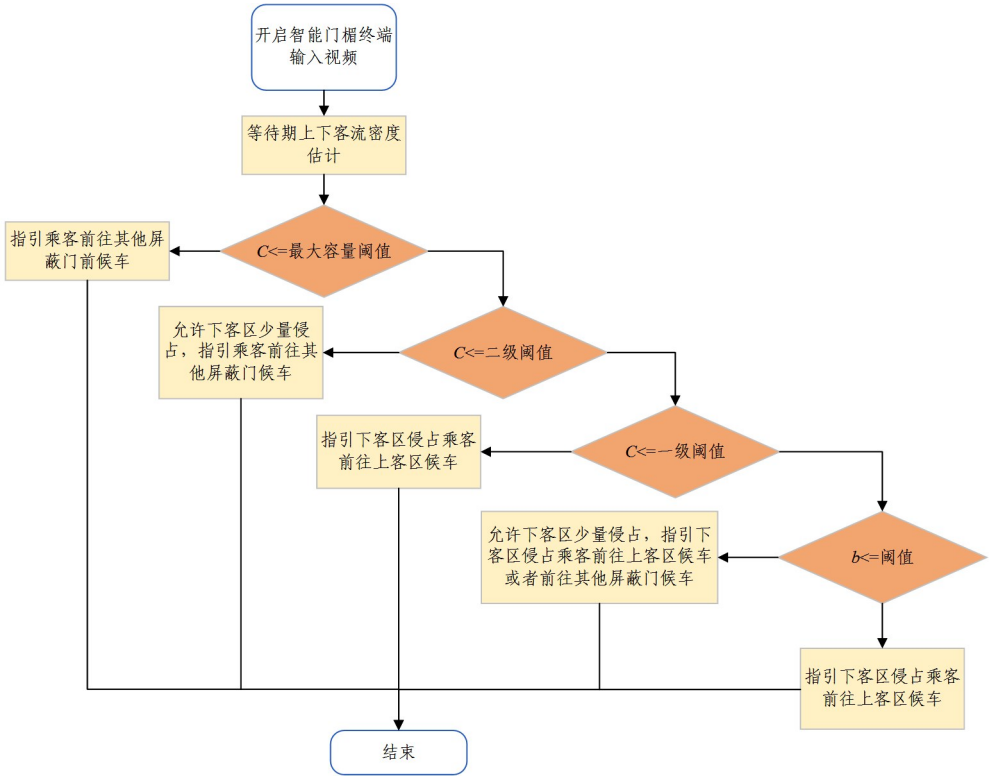


图 10 基于密度值的管控流程

Fig. 10 Control process based on density values

4 实验与分析

4.1 实验环境

本实验的基本参数以及训练的深度学习环境如表 2 所列。使用 PyTorch 深度学习框架定义网络模型，代码语言环境为 Python 3.7 版本。

表 2 实验环境配置

Table 2 Experimental environment configuration

名称	配置
操作系统	Windows 11
开发环境	CUDA 11.4
CPU	AMD Ryzen 7-5800H
显卡	NVIDIA GeForce RTX 3060 Laptop 6GB
内存	16GB
部署设备	旭日 X3 派

4.2 实验数据集

本文训练目标检测网络的数据集来自地铁站内采集的监控视频，使用 Python 导入 opencv 包对视频进行读取，并使用 5FPS 的帧率进行抽帧，生成的图片分辨率为 1920×1080。通过对不同时间段地铁视频的采集，建立了丰富的站内乘客数据，提供训练模型多样化的样本数据。数据集共包含 8738 张图片，对获取的每一帧图片使用 Labelimg 软件进行手动标注，标注界面如图 11 所示，标注的真实标签为 PASCAL VOC 格式的 XML 文件，其中包括对象边界框的坐标信息，由左上角的坐标（x_{min}，y_{max}）和右下角的坐标（x_{max}，y_{min}）组成，本文选择头部边界框的中点坐标作为密度估计中行人所在区域的判断位置。然后将 XML 文件归一化得到归一化的坐标以及归一化的宽高，转换为 YOLOv5s 算法适配的 txt 训练格式。标注目标选为行人的头部，标签为“head”类，以此来消除行人身体部分遮挡带来的影响。实验按照 8∶2 的比例划分图片，训

练集有 6166 张，验证集有 2622 张，一共包含 41925 个标签。

图 11　Labelimg 标注界面

Fig. 11　Labelimg annotation interface

本文验证密度估计效果的数据集通过搭建视频采集场景制作，本文通过模拟智能门楣终端相机的高度和角度进行拍摄。拍摄的设备为树莓派相机，拍摄帧率设置为 60 帧/秒，将拍摄的视频保存为.avi 格式，然后使用 ffmpeg 分帧工具将视频转换为图像，图像的像素为 720 * 1080，部分数据集样本如图 12 所示。

图 12　部分数据集样本

Fig. 12　Partial dataset samples

数据集中分别模拟了上下客区中不同乘客密度值的 5 种场景，分别对应 5 种不同的管控措施，用 5 种场景测试在不同密度值下本文提出的密度估计模型的性能，各场景下对应的采集数量如表 3 所列。

表 3　不同密度值下的数据分布情况

Table 3　Data distribution at different density values

不同密度值	采集数量/张
(1)	478
(2)	356
(3)	322
(4)	452
(5)	421

4.3　评价标准

首先将网络的输入图片大小统一裁剪为 640×640 的尺寸，设置 batchsize 为 4，模型初始学习率设置为 0.01，优化器选择随机梯度下降法来更新网络参数，使用余弦函数动态调整，Dropout 率设置为 0.05，在训练集上最大训练次数 epoch 设置为 200，前 3 个 epoch 用作热身训练，热身初始学习率因子为 0.001，IoU 训练时的置信度阈值设置为 0.45。

为了更好地衡量目标检测算法的精确度和实时性能，本研究设置的模型检测性能评价指标有：精确率(Precision，P)、召回率(Recall，R)、平均精度(Average Precision，AP)、浮点计算量(Giga Floating-point Operation Persecond，GFLOPs)、每秒传输帧数(Frames per Second，FPS)、参数量(Parameters)。

精确率 P 定义为在所有的检测目标中预测正确的概率，计算式为：

$$Precision=\frac{TP}{TP+FP} \tag{16}$$

召回率 R 定义为实际存在的图片中预测正确的概率，计算式为：

$$Recall=\frac{TP}{TP+FN} \tag{17}$$

平均精度 AP 定义为不同召回率下的平均精确度，计算式为：

$$AP=\int_0^1 p(R)\mathrm{d}R \tag{18}$$

其中，TP(True Positive)表示正样本中被正确检测的例子，FP(False Positives)表示负样本被预测为正样本的例子，FN(False Negatives)表示正样本被错误预测为负样本的例子，AP 表示 P(R)(precision-recall)学习的模型检测性能的好坏，GFLOPs 表示模型推理时的浮点运算次数，较高的 GFLOPs 值可能表示较大的计算量，需要更多的计算资源，Parameters 表示模型中包含参数的数量，可以用来衡量网络模型的复杂度，FPS 表示模型的检测速度，即检测图片的数量和检测时间的比例。

4.4　实验结果与分析

首先在地铁数据集上对 MCA-YOLOv5s 目标检测算法中提出的各个模块进行实验，然后与其他的主流目标检测方法进行比较，最后在拍摄的不同乘客密度值数据集上测试 MCA-YOLOv5s 算法与原始 YOLOv5s 算法的密度估计结果。MCA-YOLOv5s 模型的训练损失曲线如图 13 所示，包括训练集和验证集的定位损失(box_loss)与置信度损失(obj_loss)，训练在 100 轮左右时达到收敛，定位损失在 0.025 左右、置信度损失在 0.018 左右趋于平滑状态。

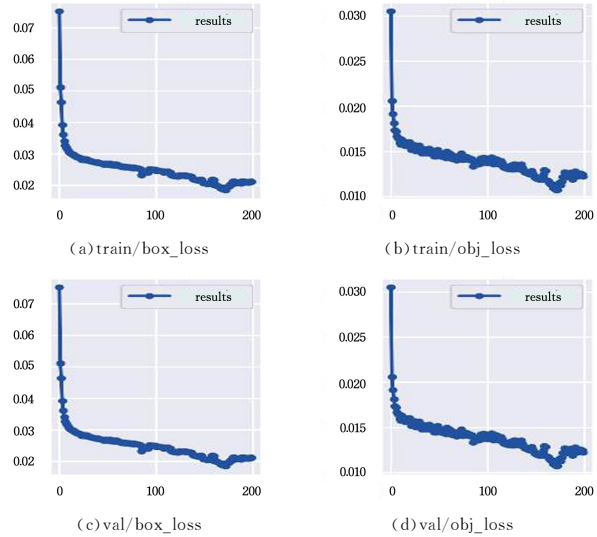


图 13　MCA-YOLOv5 训练损失曲线图

Fig. 13　Curve of MCA-YOLOv5 training loss

为了验证本文提出的轻量级主干网络 P-Mobilenetv3、融合坐标注意力机制的 C3CA 模块、去掉最后一个 20×20 的小尺寸检测层和 Alpha IoU 损失函数对模型性能的影响，首先构建 YOLOv5s 的基准网络，并在此基础上逐步加入各个提出的模块，在相同的数据集和训练环境下依次对各个模块进行实验，分析不同模块对实验结果的影响。实验结果如表 4 所列。

表 4 中采用 AP，FPS 以及 GFLOPs 3 种指标评价模型。由实验结果可知：在模型主干替换为改进的轻量化模块 P-Mobilenetv3 后模型的计算量大幅减少，FPS 提升了 66.2%，

但是检测的平均精度 AP 下降了 1.5%，说明该模块能够极大地加快网络对特征的提取速度，但是提取到的特征精度略微有所下降；在此基础上在特征融合阶段的 C3 模块中融入 CA 注意力模块，在小幅降低检测速度的同时，目标检测 AP 提高了 1.6%，弥补了轻量化结构导致的部分精度损失，说明添加该模块使得网络模型在特征融合阶段关注了目标的位置信息，有效地提高了模型的检测精度；最后再将 CIoU 损失函数替换为 Alpha IoU，目标检测的精度和速度都得到了提升，说明该损失函数能够正确地处理 High Loss 的目标，从而提高检测性能. 将所有模块都融入 YOLOv5s 模型后得到的 MCA-YOLOv5s 模型最终 AP 提高了 0.2%，FPS 加快了 59.0%，因此本文模型不仅有效地减小了模型的计算量和复杂度，加快了检测速度，同时维持了屏蔽门前对目标的检测精度。

表 4 MCA-YOLOv5s 消融实验结果

Table 4 Ablation experimental results of MCA-YOLOv5s

算法	+PMobilenetv3	+CA	-20×20head	+AlphaIoU	FLOPs	*AP*/%	FPS
YOLOv5s					15.8×10^9	94.6	43.2
A	√				4.3×10^9	93.1	71.8
B	√	√	√		3.9×10^9	94.7	68.2
C	√	√	√	√	3.9×10^9	94.8	68.7

为了验证本文提出的 P-Mobilenetv3 轻量化主干网络的有效性，将 MCA-YOLOv5s 网络中的 P-Mobilenetv3 替换为目前主流的轻量化网络 Mobilenetv3，Ghostnet 和 Fastenet，4 种网络模型使用相同的参数和预处理方式，并在相同的数据集下加载预训练权重进行对比实验，不同网络模型的性能对比如表 5 所列。与当前主流的同类型方法相比，本文提出的 P-Mobilenetv3 轻量化网络结构在 AP 和 FPS 评估指标上取得了最好的结果，综合来看 P-Mobilenetv3 轻量化网络结构最适合本文的检测任务。

表 5 不同主流轻量化网络实验结果

Table 5 Experimental results of different mainstream lightweight networks

轻量化网络	FLOPs	*P*/%	*R*/%	*AP*/%	FPS
Ghostnet	$\mathbf{7.90\times10^9}$	94.9	91.8	94.4	56.8
Mobilenetv3	$\mathbf{1.90\times10^9}$	93.8	93.5	94.3	64.6
Fasternet	11.8×10^9	93.5	**94.2**	94.6	66.1
P-Mobilenetv3	4.30×10^9	94.1	93.7	**94.8**	**68.7**

在对大尺度特征的单阶段检测网络中，YOLOv4，YOLOv5s 和 YOLOXs 是实际应用中常见的网络，YOLOv7 是近期提出的网络，Faster R-CNN 是传统的 two-stage 网络，下面将本文提出的 MCA-YOLOv5s 网络与上述所提的 5 个主流网络进行对比实验，实验结果如表 6 所列。

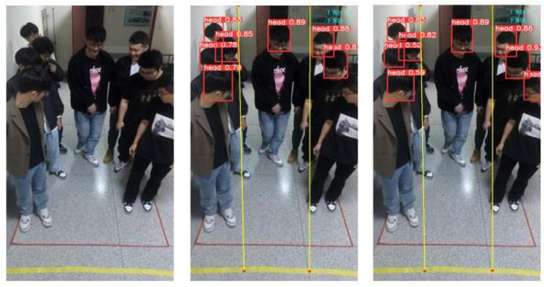
表 6 不同主流目标检测网络的实验结果

Table 6 Experimental results of different mainstream object detection networks

网络	*P*/%	*R*/%	*AP*/%	FPS
Faster R-CNN	64.5	89.4	89.4	10.9
YOLOv4	92.5	80.4	80.2	18.6
YOLOXs	87.4	92.8	93.9	38.3
YOLOv7	93.7	**93.9**	**95.2**	32.1
YOLOv5s	93.5	92.8	94.6	43.2
MCA-YOLOv5s	**94.1**	93.7	94.8	**68.7**

从表中可以看出，本文提出的网络精确率 P% 和 FPS 指标达到了最佳，分别为 94.1% 和 68.7。YOLO7 则有最高的召回率和平均精度，分别比本文网络高 0.2 和 0.4 个百分比，但是其 FPS 只有 32.1，相比于本文网络低了 36.6，并且其在模型部署的实际应用中还不成熟，文档和社区资源都不完善，难以部署在计算资源受限的设备平台。综上所述，本文提出的 MCA-YOLOv5s 网络的综合性能优于其他主流目标检测网络，适合部署在地铁内大规模的智能门楣监控系统设备中。

最后进行密度估计结果的实验，分别使用本文提出的 MCA-YOLOv5s 网络和原始的 YOLOv5s 网络进行测试，测试数据集选择拍摄的不同密度值下的乘客，实验结果如图 14 所示。本文提出的网络相比于 YOLOv5s，在极大提升检测速度的同时，对乘客的密度估计仍然具有高的准确性。



(a)原图 (b)YOLOv5 (c)MCA-YOLOv5

图 14 密度估计结果对比图

Fig. 14 Comparison chart of density estimation results

结束语 本文研发了基于目标检测的等待期上下客区客流指引方法。首先针对地铁场景中智能门楣终端相机大规模的采集需求，提出了一种轻量化的密度估计网络模型。然后根据智能门楣终端的安装高度以及真实场景中地铁屏蔽门前上客区和下客区的范围确定了相机的参数和安装的角度，确保每个乘客被正确地统计在对应的区域中，并针对密度估计结果制定了一套基于不同密度值的管控流程。最后通过实验表明，本文提出的客流指引方法能够快速且准确地获取屏蔽门前大尺度乘客的位置信息，在保证密度估计准确性的前提下极大地提高了模型的推理速度，因此能够有效地降低实际工程中成本的投入。

由于不同城市屏蔽门前场景大小不同，本文使用的基于 MCA-YOLOv5s 目标检测的密度估计方法适用于南京地铁上下客区范围较小的场景，若应用于较为密集的上下客区场景，在保证检测速度满足需求的情况下，使用基于密度图回归的方法将是更好的选择。

参 考 文 献

[1] GIRSHICK R. Fast r-cnn[C]// Proceedings of the IEEE International Conference on Computer Vision. 2015:1140-1448.

[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 and Machine Intelligence,

2016，39(6)：1137-1149.

[3] HE K，GKIOXARI G，DOLLÁR P，et al. Mask r-cnn[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017：2961-2969.

[4] LIU W，ANGUELOV D，ERHAN D，et al. Ssd：Single shot multibox detector[C]// Computer Vision-ECCV 2016：14th European Conference，Amsterdam，The Netherlands，Part I 14. Springer International Publishing，2016：21-37.

[5] REDMON J，FARHADI A. YOLO9000：better，faster，stronger [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017：7263-7271.

[6] SU P，HAN H，LIU M，et al. MOD-YOLO：Rethinking the YOLO architecture at the level of feature information and applying it to crack detection[J]. Expert Systems with Applications，2024，237：121346.

[7] GOEL L，PATEL P. Improving YOLOv 6 using advanced PSO optimizer for weight selection in lung cancer detection and classification[J]. Multimedia Tools and Applications，2024，83(32)：78059-78092.

[8] WANG C Y，BOCHKOVSKIY A，LIAO H Y M. YOLOv7：Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023：7464-7475.

[9] LIU S，QI L，QIN H，et al. Path aggregation network for instance segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018：8759-8768.

[10] HOWARD A，SANDLER M，CHU G，et al. Searching for mobilenetv3[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019：1314-1324.

[11] CHEN J，KAO S，HE H，et al. Run，Don't Walk：Chasing Higher FLOPS for Faster Neural Networks[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023：12021-12031.

[12] HOU Q，ZHOU D，FENG J. Coordinate attention for efficient mobile network design[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021：13713-13722.

[13] HE J，ERFANI S，MA X，et al. α-IoU：A family of power intersection over union losses for bounding box regression[J]. Advances in Neural Information Processing Systems，2021，34：20230-20242.

[14] HOWARD A G，ZHU M，CHEN B，et al. Mobilenets：Efficient convolutional neural networks for mobile vision applications[J]. arXiv：1704. 04861，2017.

[15] SANDLER M，HOWARD A，ZHU M，et al. Mobilenetv2：Inverted residuals and linear bottlenecks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018：4510-4520.

[16] HU J，SHEN L，SUNG. Squeeze-and-excitation networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018：7132-7141.

[17] WANG Q，WU B，ZHU P，et al. ECA-Net：Efficient channel attention for deep convolutional neural networks[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020：11534-11542.

[18] WOO S，PARK J，LEE J Y，et al. Cbam：Convolutional block attention module[C]// Proceedings of the European Conference on Computer Vision(ECCV). 2018：3-19.

[19] REZATOFIGHI H，TSOI N，GWAK J Y，et al. Generalized intersection over union：A metric and a loss for bounding box regression[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019：658-666.

[20] ZHENG Z，WANG P，REN D，et al. Enhancing geometric factors in model learning and inference for object detection and instance segmentation[J]. IEEE Transactions on Cybernetics，2021，52(8)：8574-8586.

LE Lingzhi，born in 1976，senior engineer. His main research interests include smart urban rail transit and intelligent substation.