

## 轨道交通视频中乘客口罩佩戴检测算法研究\*

李永玲<sup>1,2</sup> 秦 勇<sup>1</sup> 曹志威<sup>1,2</sup> 谢征宇<sup>2</sup> 吴志宇<sup>1,3</sup>

(1. 北京交通大学轨道交通控制与安全国家重点实验室, 100044, 北京; 2. 北京交通大学交通运输学院, 100044, 北京;

3. 北京交通大学软件学院, 100044, 北京//第一作者, 硕士研究生)

**摘 要** 为应对新冠疫情下乘客乘坐轨道交通时必须佩戴口罩的情况,提出一种基于深度学习的轻量化口罩检测算法(Mask-Det 算法)。首先,使用轻量化骨干网络 EfficientNet 提取图像特征;然后,利用高效的特征融合模块增强用于检测小目标的浅层特征图的语义信息;最后,算法在公共场景数据集上训练,并使用迁移学习在轨道交通数据集上做进一步优化。Mask-Det 算法检测准确率高、模型参数小、检测速度快,可以实时检测各场所乘客是否佩戴口罩,有效减轻工作人员压力,提高进站速度。

**关键词** 车站安全; 轨道交通视频; 口罩佩戴检测; 新冠疫情

**中图分类号** F530.7; R126.4

**DOI:**10.16037/j.1007-869x.2022.12.014

## Detection Algorithm of Passengers Wearing Masks in Rail Transit Video

LI Yongling, QIN Yong, CAO Zhiwei, XIE Zhengyu, WU Zhiyu

**Abstract** In response to the mandatory situation of passengers wearing masks on rail transit during COVID-19 pandemic period, a lightweight mask detection algorithm (Mask-Det) based on deep learning is proposed. First, the lightweight backbone network EfficientNet is used to extract image features. Then, a highly efficient feature fusion module is used to enhance the semantic information of the shallow feature map for detecting small targets. Finally, the algorithm is trained on the dataset of public scenarios, and then further optimized on the dataset of rail transit scenarios using transfer learning. Mask-Det algorithm has high detection accuracy, small model parameters, and fast detection speed. The algorithm can detect in real time whether passengers are wearing masks at various places, thus effectively alleviate personnel stress and improve passenger entry speed.

**Key words** station safety; rail transit video; mask wearing

detection; COVID-19 Pandemic

**First-author's address** State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, 100044, Beijing, China

新冠病毒(COVID-19)的爆发给社会经济带来了巨大的影响。轨道交通作为复工复产的主要交通工具,其空间密闭、人员密集且流动性广的特点有利于病毒的传播。为了降低人们在公共场所感染新冠肺炎的概率,我国疾病控制和预防中心要求乘客乘坐轨道交通(地铁、火车)时须佩戴口罩<sup>[1]</sup>。但是轨道交通防疫人员有限,因此需要口罩智能识别算法来检测安检口、闸机口及大厅等场所的乘客是否佩戴口罩,并设置自动语音提醒。这可以提高进站速度,减轻轨道交通防疫人员的工作压力<sup>[2-3]</sup>。

佩戴口罩检测研究属于人脸检测的范畴。基于深度学习的人脸检测算法表现较好<sup>[4]</sup>。文献[5]提出了一种单步多尺度目标检测器(Single Shot MultiBox Detector, 简为 SSD)<sup>[6]</sup>的口罩检测模型,其通过 K-Means 聚类的方法确定标注数据集中人脸框的长宽比分布,修改 SSD 算法的锚框(anchor)比例。该检测算法符合实时性要求,但是没有针对轨道交通场景做适配。文献[7]在 RetinaNet 模型的基础上提出了口罩检测模型,以 ResNet 为特征提取骨干网络,增加了卷积块注意力机制(Convolutional Block Attention Module, 简为 CBAM)来调整感受野的大小使其关注感兴趣的检测区域。由于深度学习算法的训练需要大量的数据集,文献[8]提出了口罩遮挡人脸检测数据集,包括模拟口罩人脸数据集和真实口罩人脸数据集。在数据集中,人脸目标较大,且不包含复杂的现实背景场景。本文提出了一种针对轨道交通场景的轻量化口罩检测

\* 中央高校基本科研业务费专项资金资助“科技领军人才团队项目”(2022JBQY007);交通运输部“交通运输行业高层次技术人才培养项目”(I18100010)

算法。该算法可以部署在不具有图形处理器 (Graphics Processing Unit, 简为 GPU) 的设备上, 实时检测乘客是否佩戴口罩。

### 1 轻量化口罩佩戴检测 Mask-Det 算法

为满足在轨道交通场景现有的中央处理器 (Central Processing Unit, 简为 CPU) 部署的要求, 本

文提出了一种轻量化口罩检测 Mask-Det 算法。图 1 为其网络结构图。该算法主要包括轻量化特征提取网络 EfficientNet-B3、轻量化特征融合模块及损失函数模块。为了提高该算法在轨道交通场景的检测效果, 本文收集整理轨道交通口罩检测数据集, 并使用迁移学习的方法提高该算法对轨道交通场景乘客是否佩戴口罩的适配性。

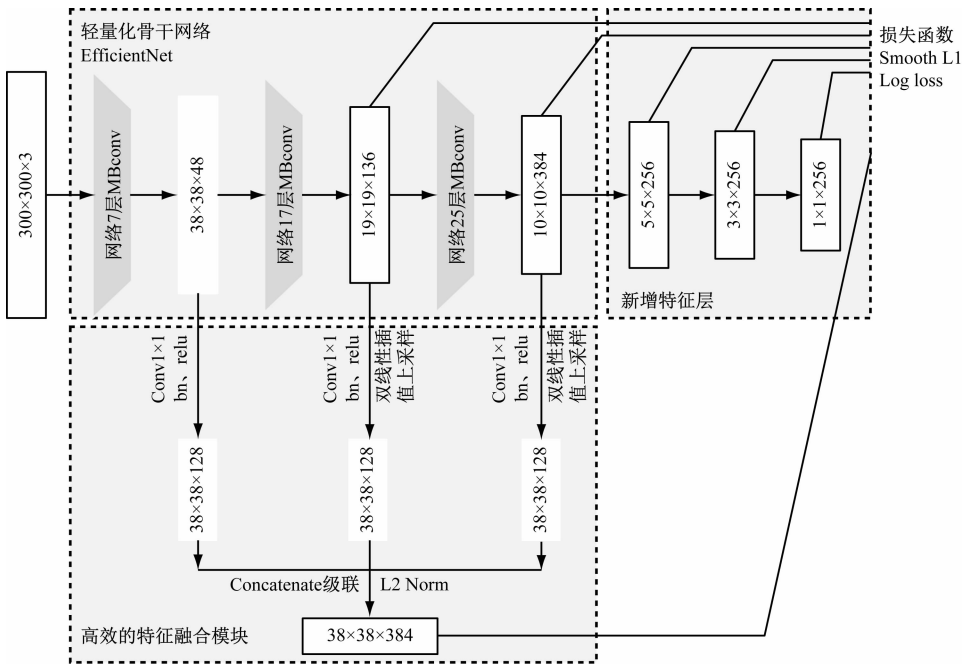


图 1 Mask-Det 算法网络结构图  
Fig. 1 Diagram of Mask-Det algorithm network structure

#### 1.1 轻量化骨干网络 EfficientNet

本文使用 EfficientNet 模型<sup>[9]</sup>作为特征提取网络模型。卷积神经网络通常采用扩展模型的深度、通道数或图像输入分辨率的方法来提高准确率。EfficientNet 模型提出了一个复合的网络缩放方法, 在节省计算资源的同时可获得更高的准确率。即:

$$\left. \begin{aligned} d &= \alpha^\phi \\ w &= \beta^\phi \\ r &= \gamma^\phi \\ \text{s. t. } \alpha\beta^2\gamma^2 &\approx 2 \end{aligned} \right\} \quad (1)$$

式中:

- $d$ ——网络深度;
- $w$ ——通道数;
- $r$ ——输入图像的分辨率;
- $\alpha$ ——分配给网络深度的计算资源参数,  $\alpha \geq 1$ ;

- $\beta$ ——分配给网络通道数的计算资源参数,  $\beta \geq 1$ ;
- $\gamma$ ——分配给图像分辨率的计算资源参数,  $\gamma \geq 1$ ;
- $\phi$ ——复合缩放系数,  $\phi = 1, 2, \dots, 7$ 。

其中:  $\alpha, \beta$  和  $\gamma$  是通过网格搜索 (Grid Search) 方法获得的常量;  $\phi$  的值越大, 需要的计算资源越多。每秒运算的浮点数 (FLOPS) 为卷积神经网络的卷积运算量, 由于卷积运算在神经网络中占主导地位, 故使用式 (1) 缩放卷积神经网络使 FLOPS 总量增加约  $(\alpha\beta^2\gamma^2)^\phi$  倍。而 EfficientNet 模型约束  $\alpha\beta^2\gamma^2 \approx 2$ , 因此, 对于任意的缩放系数  $\phi$ , 运算量为原来的  $2^\phi$  倍。EfficientNet 模型采用 MnasNet 结构<sup>[10]</sup>进行多目标神经网络结构搜索, 构建了 FLOPS 为 400 M 的 EfficientNet-B0 骨干神经网络, 其网络

结构如表 1 所示。

表 1 EfficientNet-B0 骨干神经网络

Tab. 1 EfficientNet-B0 backbone neural network				
阶段	卷积运算	分辨率	通道数	层数
1	Conv3 × 3	224 像素 × 224 像素	32	1
2	MBConv1, 3 × 3	112 像素 × 112 像素	16	1
3	MBConv6, 3 × 3	112 像素 × 112 像素	24	2
4	MBConv6, 5 × 5	56 像素 × 56 像素	40	2
5	MBConv6, 3 × 3	28 像素 × 28 像素	80	3
6	MBConv6, 5 × 5	28 像素 × 28 像素	112	3
7	MBConv6, 5 × 5	14 像素 × 14 像素	192	4
8	MBConv6, 3 × 3	7 像素 × 7 像素	320	1
9	Conv1 × 1 + Pooling + FC	7 像素 × 7 像素	1 280	1

注:1 × 1、3 × 3、5 × 5 表示卷积核大小为 1 × 1、3 × 3、5 × 5;Conv 表示卷积操作;Conv3 × 3 表示卷积核为 3 × 3 的卷积操作,Conv1 × 1、Conv5 × 5 同理;Pooling 为池化层;FC 为全连接层。

EfficientNet-B3 网络能较好地实现检测速度和准确率的平衡,因此本文选用 EfficientNet-B3 网络作为轻量化的特征提取网络模型<sup>[11]</sup>。从 EfficientNet-B0 网络得到 EfficientNet-B3 网络需要经过两个步骤:第一步,设  $\phi = 1$ ,通过网格搜索得到  $\alpha = 1.2$ ,  $\beta = 1$ ,  $\gamma = 1.15$ ;第二步,令式(1)中  $\phi = 3$ ,缩放 EfficientNet-B0 网络的深度、通道数和所输入图片的分辨率,即得到 EfficientNet-B3 网络,其结构见图 2。

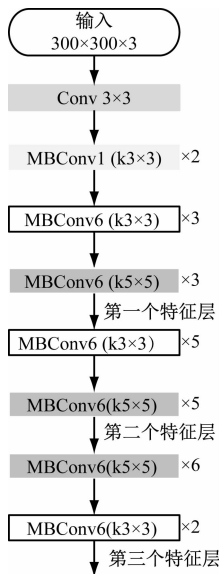


图 2 轻量化特征提取网络 EfficientNet-B3  
Fig. 2 Lightweight feature extraction network EfficientNet-B3

本文将输入的图像调整为 300 像素 × 300 像素,从 EfficientNet-B3 中提取第 7、17、25 层特征图,

并在第 25 层特征图的基础上新增三层特征图。图 1 所示上述 6 层特征图的大小依次为 38 像素 × 38 像素、19 像素 × 19 像素、10 像素 × 10 像素、5 像素 × 5 像素、3 像素 × 3 像素和 1 像素 × 1 像素。为了与最终添加特征融合的 Mask-Det 网络对比,称该网络为 Efficient-Mask 网络。

1.2 高效的特征融合模块

浅层特征图包含更多的位置、细节信息,适用于检测小目标;但由于经过的卷积运算次数少,用于识别的语义信息不够丰富<sup>[12]</sup>。本文提出一种轻量化的特征融合方法,为浅层特征图融合高层特征图的语义信息,从而在没有降低算法实时性的同时,提高算法对于小人脸的检测准确率。

对 38 × 38、19 × 19、10 × 10 的 3 层特征图的信息进行融合。如图 1 所示,首先,采用 1 × 1 的卷积将上述 3 层特征图的通道数变为 128;然后,对 19 × 19、10 × 10 的 2 张特征图做双线性差值上采样,得到 3 张相同维度的特征图;接着,对这 3 张特征图采用 concatenate 级联;最后,引入 L2Norm 算法<sup>[6]</sup>归一化将级联后的特征图中每个位置的特征范数缩放到 5。最终可得融合了高层语义信息和低层局部信息的 38 × 38 特征图,提高了算法对小目标人脸的检测效果。

1.3 损失函数

Mask-Det 口罩检测算法的损失函数是分类损失  $L_{\text{conf}}$  和定位损失  $L_{\text{loc}}$  的加权,其中,  $L_{\text{conf}}$  是多个类别分类置信度  $c$  上的 softmax 损失<sup>[6]</sup>,  $L_{\text{loc}}$  为预测值  $l$  和真实值  $g$  之间的 Smooth L1 损失。即:

$$L(x, c, l, g) = \frac{1}{N} (L_{\text{conf}}(x, c) + \theta L_{\text{loc}}(x, l, g))^{[6]} \tag{2}$$

式中:

- $N$ ——匹配的默认框数量,如果  $N = 0$ ,则  $L(x, c, l, g) = 0$ ;
- $x$ ——指示参数; $x = 1$  表示  $t$  与  $g$  匹配, $x = 0$  表示  $t$  与  $g$  不匹配;
- $\theta$ ——定位损失  $L_{\text{loc}}$  的权重。

1.4 口罩佩戴检测数据集

基于深度学习的人脸检测算法需要大量的数据样本来进行训练。考虑到轨道交通运营的特殊性,难以获取大量的训练图像,故本文分别创建了公共场景和轨道交通场景的口罩检测数据集。2 个数据集的示例图如表 2 所示。

表 2 公开场景和轨道交通场景的口罩检测数据集示例图

Tab.2 Example pictures of mask detection datasets for public and rail transit scenarios

场景	未佩戴口罩	佩戴口罩	有遮挡	不规范佩戴口罩
公开场景				
轨道交通场景				

公共场景的口罩检测数据集由 WIDER Face<sup>[13]</sup>数据集、MAFA<sup>[14]</sup>遮挡人脸数据集及公共场景佩戴口罩人脸数据集组成,其中训练集包含 22 631 张图像,测试集包含 3 000 张图像。轨道交通场景的口罩检测数据集包含 2 500 张图像,其中训练集和测试集的图像数分别为 2 200 张与 300 张。

2 试验

2.1 试验环境与方法

本文试验采用了深度学习框架 Pytorch<sup>[15]</sup>, Ubuntu16.04 系统,训练所用 GPU(图形处理器)型号为 NVIDIA P100。测试所用设备为 Jeson Nano 嵌入式设备、Intel Core i5-6300HQ CPU 笔记本、Intel Core i7-8700K CPU 台式机。

训练过程中使用的超参数为:批处理大小(batch size)为 16,动量为 0.9,初始化学率为 0.001;经过 42 轮(epoch)和 52 轮公共场景数据集训练,学习率分别下降为 0.000 1、0.000 01。整个训练过程包含 64 轮(大约 16 万次迭代)。Mask-Det 算法在轨道交通数据集上迁移学习 4 万次迭代,学习率为 0.000 01。

为了验证模型的有效性,本文提出的 Mask-Det 算法将同 Faster R-CNN<sup>[16]</sup>、YOLOv3<sup>[17]</sup>、SSD<sup>[6]</sup>等算法进行比较。首先,在公共场景的口罩检测数据集上进行各算法的训练;然后,采用迁移学习的方法,在轨道交通场景的数据集上继续训练 Mask-Det 算法,增加网络对轨道交通场景的适配性。

2.2 评价指标

本文采用目标检测算法常用评价指标,即平均

准确率  $R_{AP}$ 、类别平均准确率  $R_{mAP}$ 、刷新帧率  $R_{FPS}$  及模型大小(单位 Mbit)。

$$p = t/(t + f) \tag{3}$$

$$r_e = t/(t + n) \tag{4}$$

$$R_{AP} = \int_0^1 p(r) dr \tag{5}$$

$$R_{mAP} = \left[ \sum_{c=1}^C R_{AP} \right] / C \tag{6}$$

式中:

- $t$ ——真正例,即正确检测到的目标;
- $f$ ——假正例,即误报的目标;
- $n$ ——假负例,即未检测到目标;
- $p$ ——查准率;
- $r_e$ ——查全率;
- $C$ ——目标检测的类别数;
- $R_{AP}$ ——平均准确率;
- $R_{mAP}$ ——类别平均准确率。

$R_{FPS}$  越大代表算法的实时性越高。模型越小代表算法越轻量化,越容易部署到现有的 CPU 设备上。

在实际应用中,准确率的定义为正确检测到的目标占测试集中所有真实值的比例。式(8)计算的准确率  $r_{准确率}$  值与式(5)计算的召回率值相等。

$$r_{准确率} = t/g \tag{7}$$

2.3 试验结果

轨道交通场景的测试集包含 300 张图像,共计 1 027 个佩戴口罩的人脸目标和 327 个未戴口罩的人脸目标,涵盖了进站口、安检口、大厅、电梯、站台、车厢等各场景的乘客。表 3 对比了本文提出的 Efficient-Mask、Mask-Det 算法与 Faster R-CNN<sup>[16]</sup>、YOLOv3<sup>[18]</sup>、SSD<sup>[6]</sup>等主流目标检测算法在轨道交

通数据集上的表现。测试设备均为 Intel Core i5-6300HQ CPU 笔记本。

由表 3 可看出:输入图像的分辨率越高、骨干网络模型越大,则检测的精度越高,但计算速度越慢;Faster R-CNN、YOLOv3 及 SSD 等算法的输入图像

分辨率分别为 600 像素×1 000像素、416 像素×416 像素、300 像素×300 像素,均大于等于 Mask-Det 算法的分辨率输入;但是 Mask-Det 的类别平均准确率为 72.77%,明显高于 Faster R-CNN 和 SSD 算法。

表 3 不同算法在轨道交通测试集上的结果对比

Tab.3 Results comparison of different algorithms in rail transit test set

算法	骨干网络	输入图像	人脸平均 准确率/%	口罩平均 准确率/%	$R_{mAP}/\%$	$R_{FPS}/(\text{帧}/\text{s})$	模型/Mbit
Faster R-CNN <sup>[16]</sup>	ResNet-50	600 像素×1 000 像素	48.10	61.38	54.74	0.1	113.9
YOLOv3 <sup>[17]</sup>	DarkNet-53	416 像素×416 像素	62.50	81.20	71.85	3.0	240.1
SSD <sup>[6]</sup>	VGG16	300 像素×300 像素	57.70	68.97	63.34	10.0	182.4
Efficient-Mask	EfficientNet-B3	300 像素×300 像素	49.03	68.45	58.74	24.0	88.5
Mask-Det	EfficientNet-B3	300 像素×300 像素	64.13	81.41	72.77	21.0	90.8

骨干网络用来提取输入图像的特征。表 3 中只有 EfficientNet-B3 是轻量化的骨干网络,所以采用该骨干网络的 Mask-Det 和 Efficient-Mask 模型较小、检测速度较快。Mask-Det 的模型大小约为 Faster R-CNN、YOLOv3、SSD 算法模型的 0.80 倍、0.38倍、0.50 倍;每秒处理的帧数是上述三种算法的 210 倍、7 倍、2.1 倍。Mask-Det 算法在普通 CPU 上每秒处理 21 帧图像,满足乘客佩戴口罩检测的实时性要求。

表 4 为 Mask-Det 算法在轨道交通数据集、不同 CPU 设备上准确率和速度的测试结果。Mask-Det 的检测准确率可以达到 96.68%,即在 300 张轨道交通测试图像(包含 1 354 个目标(all ground truths))中,能正确识别到 1 309 个目标。Mask-Det 算法在嵌入式设备 Jeson Nano、Intel Core i5-6300HQ CPU 及 Intel Core i7-8700K CPU 上的处理速度分别为 7 帧/s、21 帧/s、61 帧/s。需要强调的是,由于采用相同的训练权重,算法的准确率和运行设备无关,所以准确率相同,速度不同。用户可以按照进站口、安检口及电梯等位置的客流大小及乘客通行时间的长短来选择不同的设备,以检测乘客是否佩戴口罩。

表 4 Mask-Det 算法在不同 CPU 设备上的测试结果

Tab.4 Test results of Mask-Det algorithm on different CPU devices

CPU 设备	$R_{FPS}/(\text{帧}/\text{s})$
Jeson Nano	7
Intel Core i5-6300HQ	21
Intel Core i7-8700K	61

























## 2.4 各算法比较

各算法的检测结果对比如表 5 所示。Faster R-CNN 算法是典型的分类和检测分开的二阶段算法。因此,与其它一阶段算法相比,Faster R-CNN 算法的检测速度最慢,无法在普通的 CPU 设备上实现实时检测。此外,Faster R-CNN 算法误报率高,可能误将耳朵或手识别为人脸,导致其平均类别准确率值较低。虽然 YOLOv3 算法对人脸检测的类别准确率只略低于 Mask-Det 算法,但该算法需先将输入图像的分辨率调整为 416 像素×416 像素,再送入 DarkNet-53 特征提取网络,而其模型的参数量为 240.1 Mbit,且检测速度仅为 3 帧/s,故该算法即使部署到轨道交通场景的现有设备上也无法满足实时检测的要求。

目前,许多研究通过单独增加卷积神经网络的  $d$ 、 $w$  及  $r$  来优化算法。EfficientNet 建立了 3 个维度之间的缩放关系,较好地达到了准确率和检测速度的平衡。本文提出的 Efficient-Mask 算法采用 EfficientNet 轻量化骨干网络及 300 像素×300 像素的网络图像输入分辨率,在减少参数量的同时使速度也最快。卷积神经网络因其浅层特征图包含更多的空间细节特征,故常用于检测小目标,但其经过的卷积运算次数少、语义信息少,导致小目标人脸的检测准确率较低。为了解决该问题,本文提出 Mask-Det 算法,在 Efficient-Mask 算法的基础上增加了特征融合模块,将深层特征图的语义信息融合到用于检测小目标的浅层特征图上。特征融合模块增加了计算量,所以与 Efficient-Mask 算法相比,Mask-Det算法的模型参数量有少量增加,算法速度

表 5 不同算法的检测对比图

Tab. 5 Detection comparison of different algorithms

场景	原图	Faster RCNN 算法	YOLOv3 算法	SSD 算法	Efficient-Mask 算法	Mask-Det 算法
场景一						
场景二						
场景三						
场景四						

有所降低,但是类别平均准确率显著提高,满足部署到现场的轻量化和实时性、高准确率的要求。

由表 5 可见:Mask-Det 算法的人脸识别框最全,说明其对小目标人脸的漏检率低,检测效果优于其他算法。此外,表 5 中每张人脸的识别框内均有相应算法对检测到目标的置信度数字评分,其中 Mask-Det 算法的评分最高。

最后,通过测试 Mask-Det 算法在 3 种不同设备上的速度和准确率,确定 Mask-Det 算法能满足轨道交通辅助不同场景工作人员检测的要求,可部署性强。

3 结语

本文提出了一种基于轨道交通监控视频的轻量化乘客佩戴口罩检测算法。首先,采用轻量化骨干网络 EfficientNet 提取特征图;然后,将深层特征图的语义信息融合到用于检测小目标的浅层特征图上,提高了该算法对小目标人脸的检测效果;最后,将该算法先后在整理的公共场景和轨道交通场景的数据集训练,提高了对轨道交通场景的适配性。

相比其他主流算法,本文提出的 Mask-Det 算法检测准确率高(类别平均准确率为 72.77%、准确率达 96.68%)、模型参数小(仅为 90.8 Mbit)、检测速度快(61 帧/s),能实时检测轨道交通安检口、闸机口、大厅等场所监控视频中的乘客是否佩戴口罩,有利于减少人员工作量、实现安防监控智能化,从而提高进站速度。

参考文献

[1] 谢征宇,曹志威,李永玲,等.基于视频的轨道交通车站乘客口罩佩戴检测及测温技术[J].中国铁路,2020(3):126.  
XIE Zhengyu, CAO Zhiwei, LI Yongling, et al. Video-based mask-wearing detection and temperature measurement technology in rail transit stations[J]. China Railway, 2020(3):126.

[2] 刘斌,丁波,赵萌萌,等.新型冠状病毒肺炎疫情防控期间武汉地铁客运管理措施分析[J].城市轨道交通研究,2020(10):5.  
LIU Bin, DING Bo, ZHAO Mengmeng, et al. Analysis of Wuhan Metro passenger transport management measures during the period of COVID-19 prevention and control[J]. Urban Mass Transit, 2020(10):5.

[3] 孙章.新冠肺炎疫情防控与新型城市轨道交通系统开发[J].城市轨道交通研究,2020(3):彩8.  
SUN Zhang. Prevention and control of novel coronavirus pneumonia and the development of new urban rail transit system[J]. Urban Mass Transit, 2020(3):C8.

[4] 徐首峰.人脸识别技术在上海城市轨道交通中的应用[J].城市轨道交通研究,2020(增刊2):164.  
XU Shoufeng. Application of face recognition technology in Shanghai urban rail transit[J]. Urban Mass Transit, 2020(S2):164.

[5] AIZOOTech. Detect faces and determine whether people are wearing mask[Z/OL]. (2020-02-18)[2021-03-24]. <https://github.com/AIZOOTech/FaceMaskDetection>.

[6] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot multibox detector[C]//Computer Vision-European Conference on Computer Vision (ECCV). Amsterdam: ECCV, 2016: 21-37.