Multimedia Tools and Applications (2022) 81:42433–42456
https://doi.org/10.1007/s11042-022-13491-x

1220: VISUAL AND SENSORY DATA PROCESSING FOR REAL TIME INTELLIGENT SURVEILLANCE SYSTEM

Face detection for rail transit passengers based on single shot detector and active learning

Zhiwei Cao[1,2] · Yong Qin[1,3] · Yongling Li[1,2] · Zhengyu Xie[2] · Jianyuan Guo[2] · Limin Jia[1,3]

Received: 12 January 2021 / Revised: 20 October 2021 / Accepted: 13 July 2022 /
Published online: 30 August 2022
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract

COVID-19 spreads rapidly among people, so that more and more people are wearing masks in rail transit stations. However, the current face detection algorithms cannot distinguish between a face wearing a mask and a face not wearing a mask. This paper proposes a face detection algorithm based on single shot detector and active learning in rail transit surveillance, effectively detecting faces and faces wearing masks. Firstly, we propose a real-time face detection algorithm based on single shot detector, which improves the accuracy by optimizing backbone network, feature pyramid network, spatial attention module, and loss function. Subsequently, this paper proposes a semi-supervised active learning method to select valuable samples from video surveillance of rail transit to retrain the face detection algorithm, which improves the generalization of the algorithm in rail transit and reduces the time to label samples. Extensive experimental results demonstrate that the proposed method achieves significant performance over the state-of-the-art algorithms on rail transit dataset. The proposed algorithm has a wide range of applications in rail transit stations, including passenger flow statistics, epidemiological analysis, and reminders of passenger who do not wear masks. Simultaneously, our algorithm does not collect and store face information of passengers, which effectively protects the privacy of passengers.

Keywords Face detection · Mask detection · Rail transit passengers · Single shot detector · Active learning

✉ Yong Qin
yqin@bjtu.edu.cn

[1] State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, No.3 Shangyuancun, Beijing 100044, People's Republic of China

[2] School of Traffic and Transportation, Beijing Jiaotong University, Beijing 100044, China

[3] Beijing Research Center of Urban Traffic Information Sensing and Service Technologies, Beijing 100044, China

1 Introduction

With the spread of COVID-19, more and more people wear mask to reduce the probability of being infected. Rail transit stations with limited space are very conducive to the spread of viruses, so several governments and companies have made mandatory measures for passengers to wear masks. With the rapid development of artificial intelligence, cameras are playing an increasingly important role in rail transit stations. Cameras are widely deployed in the station entrance and have accumulated a large amount of video, which provide a solution for detecting whether passengers wear masks. The cameras installed at the entrance of rail transit station are shown in Fig. 1. The proposed algorithm can recognize whether passengers wear masks, which not only counts the data of passengers wearing masks for epidemic analysis, but also helps the station remind passengers to wear masks by using the microphones.

After the outbreak of COVID-19, several face and mask detection algorithms [4, 13, 24, 27, 28, 38, 41, 42, 45] are studied. [4, 13, 24, 28, 41] propose the fast algorithms to detect face and mask, but they have some false positives and false negatives. Several face mask detection algorithms [27, 38, 42] are proposed to improve accuracy and have good results, but their speed is slow. It is difficult to improve accuracy while ensuring real-time processing. Moreover, as far as we know, there is no research on the mask detection algorithm applied to rail transit. In the rail transit, face sizes are small and face angles are diverse. There is a challenge with these algorithms for mask detection in rail transit: these algorithms are trained on datasets lacking rail transit samples, so they may have poor generalization in rail transit scenarios. Therefore, it is necessary to make a mask dataset of rail transit. Although there are many images of face and mask from rail transit, two problems still exist when applied in rail transit. First, the labeling of sample data takes a lot of time and automatic labeling causes many errors. Second, the sample data has a high degree of similarity, so the marginal benefit generated in training is very small. A lot of similar samples cause sample imbalance, making it more difficult to detect minority objects.

To solve these problems, we propose a face detection algorithm based on single shot detector (SSD) called SSD-Mask and a novel active learning for rail transit passengers. The proposed algorithm improves the accuracy to detect face and face wearing a mask, and utilizes

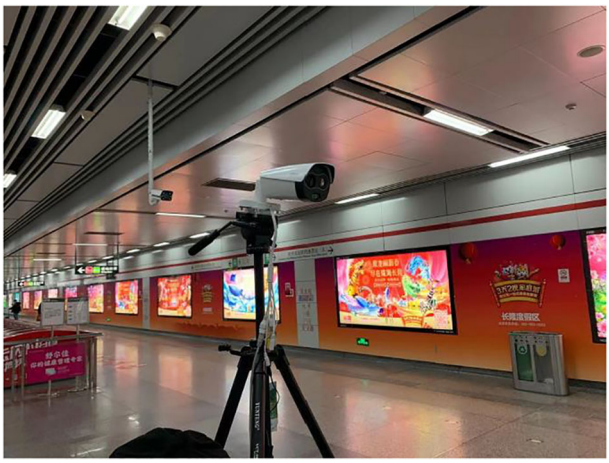

Fig. 1 The cameras installed at the rail transit station

a novel active learning method to obtain valuable samples of rail transit, so as to train a face detection model suitable for rail transit scenarios.

The contributions of this paper are as follows:

Firstly, SSD-Mask is proposed to improve accuracy while ensuring real-time processing, which optimizes the anchor aspect ratios, activation function, and loss function, and adds feature pyramid network and spatial attention module.

Secondly, a novel active learning method called semi-supervised active learning (SSAL) is proposed for screening valuable samples of rail transit and reducing sample labeling time. SSAL reduces the cost of model transfer and application, and improves the generalization of the model.

Finally, we make a high-quality mask dataset to fine-tuning the proposed algorithm. This mask dataset covers people wearing masks and face images in rail transit scenarios.

The rest of the paper is organized as follows. In the next section, related work is introduced in detail, which includes face mask detection, SSD, and active learning. In Section 3, SSD-Mask, and SSAL are presented. Section 4 shows the experimental results and discussion in detail. The conclusion is drawn in Section 5.

2 Related work

In this section, we introduce the work related to this paper from three aspects: face mask detection, SSD, and active learning.

2.1 Face mask detection

With more and more people wear mask, recent related works [4, 13, 24, 27, 28, 38, 41, 42, 45] that do a face detection to detect people wear mask or not. AIZOO [4] proposes a lightweight mask detection network based on SSD [23] with fast speed. [13] presents a mask detection network based on RetinaNet [22], which uses feature pyramid network to fuse high-level information and context attention module to focus on faces and masks. [28] proposes a lightweight mask detection algorithm based on SSD [23] and MobileNetV2 [37]. Based on the YOLOv2 [33], Loey et al. [24] uses ResNet50 as the backbone network to detect faces wearing masks in images. [41] applies YOLOv3 directly to face and mask detection, and compares it with Faster R-CNN [35]. [4, 13, 24, 28, 41] have fast speed, but their accuracy needs to be further improved. Sethi et al. [38] integrates SSD [23] and Faster R-CNN [35] for object detection, and uses transfer learning to fuse high-level semantic information in the feature map to improve the accuracy of the mask detection. [27] proposes a hybrid deep transfer learning model with machine learning method for mask detection, which uses deep learning to extract features and multiple machine learning methods to classify them. Baidu proposes a mask detection algorithm based on PyramidBox [42], which improves accuracy by using low-level feature pyramid network, context-sensitive predict module, and pyramid anchors. [27, 38, 42] have high accuracy, but their speed is slow. Therefore, a mask detection algorithm should be studied which has high accuracy and fast speed.

Although some companies have collected a large number of images of faces and masks, these data have not been made public. Several researchers synthesize and publishes the mask dataset. AIZOO [4] publishes a real mask dataset for the first time, and it promotes the development of mask detection. Wang et al. [45] make and test a mask dataset composed of

real-world images and synthesized images. In the rail transit, face sizes are small and face angles are diverse. Mask detection model trained on other scene image may have poor generalization in rail transit scenarios. Therefore, it is necessary to make a mask dataset of rail transit. Although there are many images of face and mask from rail transit, two problems still exist when applied in rail transit. First, manual labeling takes a lot of time and automatic labeling causes many errors. Second, the sample data has a high degree of similarity, so the marginal benefit generated in training is very small. A lot of similar samples also cause sample imbalance, making it more difficult to detect minority objects.

To solve these problems, we propose SSD-Mask and SSAL for rail transit passengers. To improve accuracy while ensuring real-time processing, the anchor aspect ratios, activation function, and loss function are optimized, and feature pyramid network and spatial attention module are added. Then, SSAL is proposed for screening valuable samples of rail transit and reducing sample labeling time.

2.2 SSD

SSD [23] is proposed for object detection with fast speed and high accuracy in 2016. As shown in Fig. 2, SSD [23] sets several anchors based on different feature maps and combines predictions from multiple feature maps with different resolutions to handle objects with different aspect ratios and sizes. Backbone network is used to extract features of multiple convolutional layers and obtains multiple feature maps. The backbone network of SSD [23] is based on VGGNet, which consists of part of VGG16 [40] and extra feature layers. The extra feature layers build conv8_2, conv9_2, conv10_2 and conv11_2 based on conv7 of VGG16 [40]. When the input resolution is 300*300, SSD [23] locates and classifies the output of the backbone network that includes conv4_3, conv7, conv8_2, conv9_2, conv10_2 and conv11_2. Finally, the non-maximum suppression algorithm is used to remove redundant candidate boxes and obtain the best detection boxes. A series of improved algorithms [12, 18, 47] based on SSD [23] are proposed for object detection and achieved good results. In addition, SSD [23] is very suitable for face detection because it can obtain multiple feature maps. Many face detection algorithms [8, 19, 48, 49] are proposed based on SSD, and they have made great progress.

2.3 Active learning

With the advent of the era of big data, we have tremendous amounts of data, but the quality and labeling of the data have become a problem. A large amount of data has

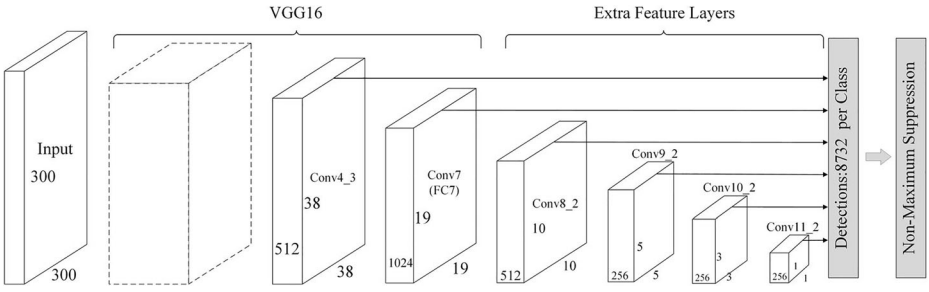

Fig. 2 SSD model framework

little effective information and is labeled for a long time, which increases the training time and does little help to the model. Active learning solves this problem by selecting the valuable samples and labeling them by experts, so few samples can be trained to get a good model. Active learning hypothesizes if a learning algorithm is allowed to choose its curious data, it will perform better with fewer data [39]. There are three scenarios in active learning: membership query synthesis, stream-based selective sampling, and pool-based active learning [39]. In this paper, we use pool-based active learning as shown in Fig. 3. Therefore, active learning below refers to pool-based active learning. Active learning sets up query strategies and utilizes the trained model to process unlabeled samples, thereby selecting valuable unlabeled samples and labeling these samples by experts. Subsequently, the labeled samples are used to retrain the model to improve the accuracy. Active learning is often used in image classification [9, 14, 16, 17, 25, 44], because the labeling of image classification task is simple and costs less time for experts. In recent years, several researches apply active learning to object detection [2, 5, 15, 30–32, 36, 43], which prove that active learning can effectively improve the accuracy of object detection. However, it is not solved that active learning requires a lot of time for labeling.

3 Methods

In this section, we introduce SSD-Mask, SSAL and mask dataset. SSD-Mask is the basis of the proposed method and is designed for face and mask detection. SSAL selects valuable rail transit images for fine-tuning SSD-Mask, which improves the robustness of SSD-Mask in rail transit scenarios. In addition, the mask dataset is made to train the proposed algorithm.

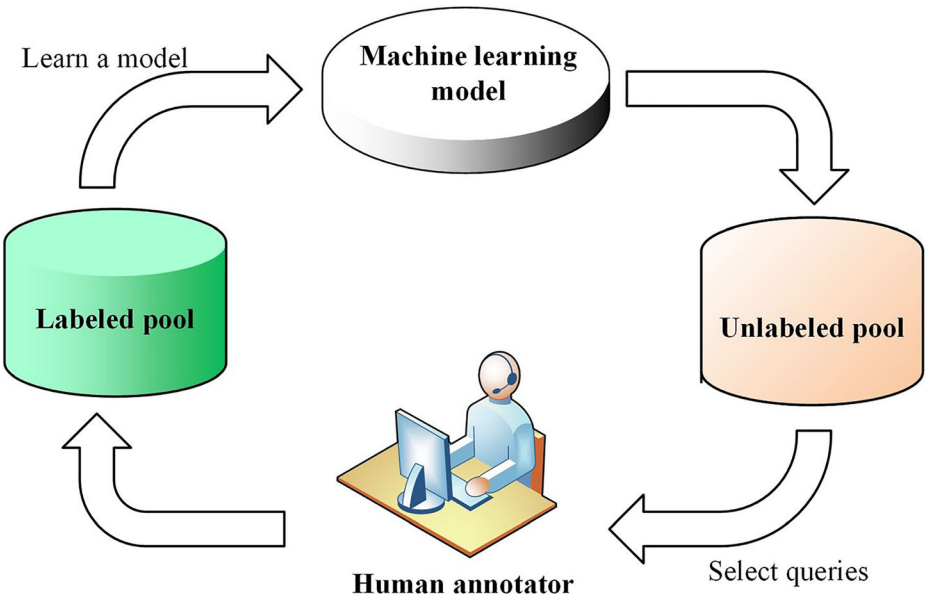

Fig. 3 The pool-based active learning

3.1 SSD-mask

To detect whether passengers wear masks, we propose a face detection algorithm called SSD-Mask based on SSD [23]. The architecture of SSD-Mask is shown in Fig. 4, including backbone network, feature pyramid network (FPN), spatial attention module (SAM), and smooth and focal loss.

As shown in Fig. 4, backbone network is the principal part of the neural network, which is used to extract the feature information of the image. FPN and SAM are used to enhance the feature information extracted by the backbone network. FPN enlarges the feature map of low-level layers while ensuring that high-level semantic information is not lost, which is conducive to detecting small objects. SAM produces a more distinguishable feature representation by selecting the focus position. Smooth and focal loss can effectively solve the problem of imbalance between positive and negative samples, thereby training a model with higher detection accuracy.

Compared with other methods in architecture, we don't change the backbone of SSD because it is efficient, but focus on enhancing the features extracted from the backbone. Specifically, [4, 24, 28] directly use features extracted from the backbone network to locate and classify. [13, 41] use FPN to enhance the high-level features extracted from backbone networks. However, since faces are smaller and low-level layers contain more effective information, we choose the low-level layers convolution layers to construct FPN. Moreover,

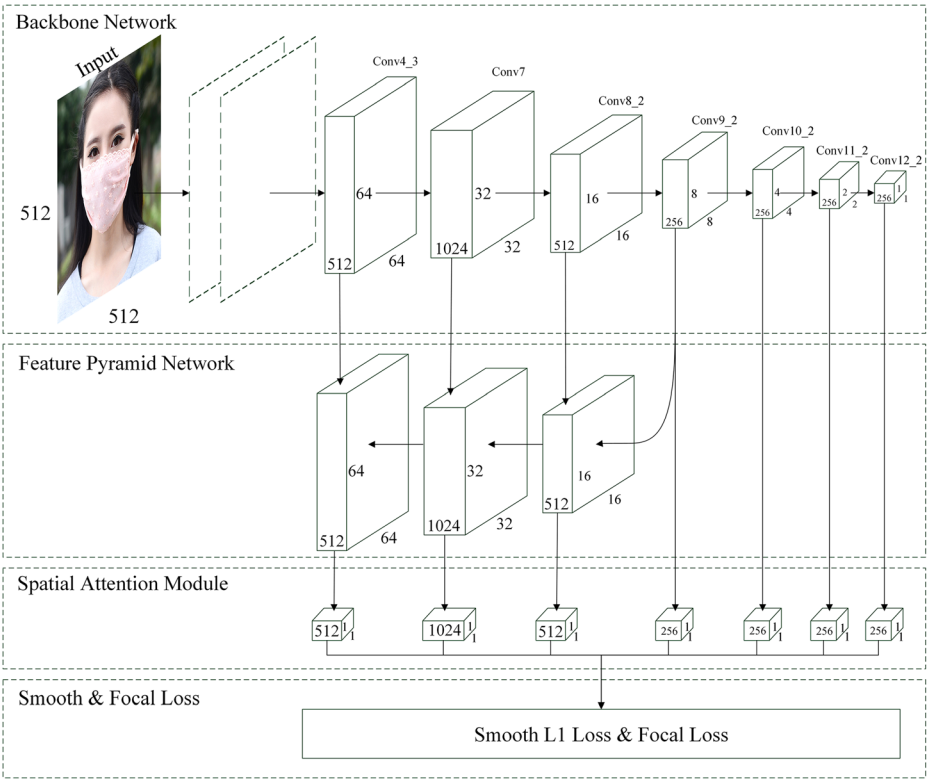

Fig. 4 Architecture of SSD-Mask

SAM is used to obtain more distinguishable feature representation. [38] combines the results of SSD [23] and Faster R-CNN [35] without modifying the framework for both.

3.1.1 Backbone network

As shown in Fig. 4, the backbone network of SSD-Mask is VGGNet [23], which consists of VGG16 [40] and extra feature layers. The backbone network uses Conv4_3 and Conv7 of VGG16 [40] to build the extra feature layers with Conv8_2, Conv9_2, Conv10_2, Conv11_2 and Conv12_2. The backbone network has a small model and fast calculations, but the effect is good.

To improve the performance of our network, the backbone network needs to match the dataset. The objects of different datasets have different aspect ratios, so we count the face aspect ratio of the dataset to make SSD-Mask match the data. There are 144,235 faces in the dataset, and the aspect ratios distribution of these faces is shown in Fig. 5. The aspect ratio of the face is mainly concentrated around 2: 1 and 1: 1. According to our experiments, the optimal aspect ratios of anchor are set.

Activation function plays an essential role in the backbone network. Mish [26] is a novel smooth activation function as shown in Fig. 6, and the research [26] demonstrates that Mish [26] is superior to Rectified Linear Unit (ReLU) [7, 11, 29] in most models. As shown in Fig. 6, Mish [26] is smoother than ReLU [7, 11, 29], which is the reason why Mish [26] is better than ReLU [7, 11, 29].To improve the performance of the model, Mish [26] is used to instead of ReLU [7, 11, 29] in the backbone network. Mish [26] is defined as follows:

$$f(x) = x \cdot tanh(ln(1 + e^x)) \tag{1}$$

3.1.2 Low-level feature pyramid network

Feature pyramid network (FPN) can make full use of the multi-scale features output by the backbone network. FPN is proposed by [21], which can detect object of different sizes to

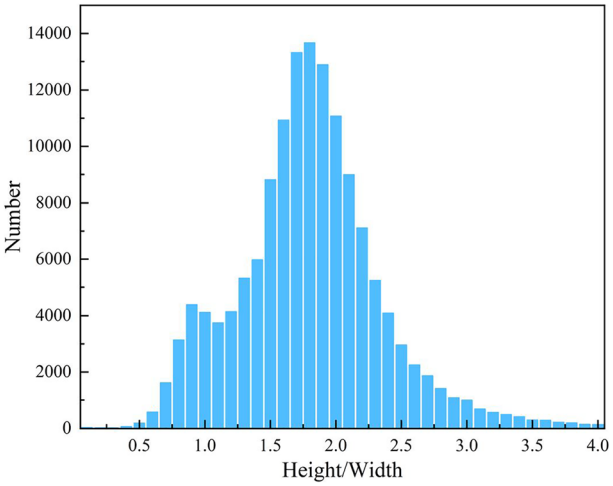

Fig. 5 Aspect ratios distribution of faces from mask dataset

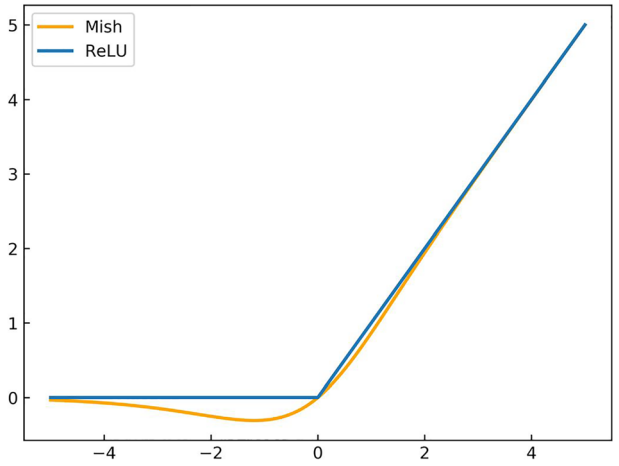

Fig. 6 Activation function: Mish and ReLU

improve the performance of the model [1, 34]. We pay more attention to the low-dimensional information of SSD-Mask, and improve the accuracy of small faces by fusing low-dimensional information. Inspired by [21, 42], we construct a low-level feature pyramid network for the output of Conv4_3, Conv7, Conv8_2, and Conv9_2 in Fig. 7.

3.1.3 Spatial attention module

Attention mechanism has been widely used in object detection, which can effectively improve algorithm performance. Although the attention mechanisms such as SE [10], CBAM [46], and SK [20] enhance the accuracy of the algorithm, the inference time increases. Spatial attention module (SAM) adjusts each position of the feature map to make the model focus on areas that deserve more attention. SAM improves network performance with very little additional inference time. As shown in Fig. 4, inspired by [1], we use SAM to deal with the output from the backbone network and the feature pyramid network. SAM is denoted as

$$\begin{aligned} x^{'} &= Conv(x) \\ y &= Sigmoid\left(x^{'}\right) \times x \end{aligned} \tag{2}$$

where $\boldsymbol{Conv}(x)$ adopts pointwise convolution for input x, that is, 1*1 convolution without changing the number of channels, and $\boldsymbol{Sigmoid}(x')$ is $1/1 + e^{-x'}$. Figure 8 shows how the spatial attention module works.

3.1.4 Smooth and focal loss

The loss of SSD-Mask consists of localization loss and confidence loss:

$$L = \frac{1}{N}\left(L_{loc} + L_{\boldsymbol{conf}}\right) \tag{3}$$

where N is the number of matched default boxes. The localization loss based on Smooth L1 loss [6] is as follows:

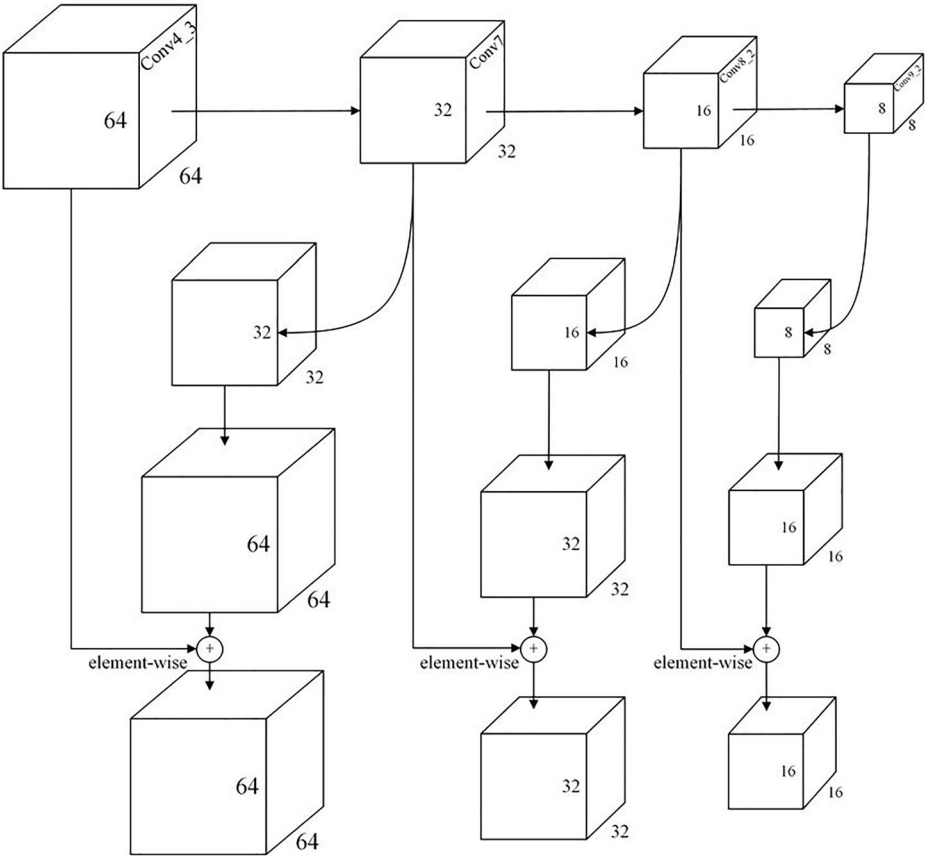

Fig. 7 Low-level feature pyramid network

$$L_{loc} = \sum_{i \in Pos}^{N} \sum_{m \in \{cx,cy,w,h\}} x_{ij}^{k} \cdot \boldsymbol{Smooth}_{L1}\left(l_i^m - \widehat{g}_j^m\right) \tag{4}$$

where i is the index of predicted positive box, j is the index of ground truth box, and x_{ij}^{k} is

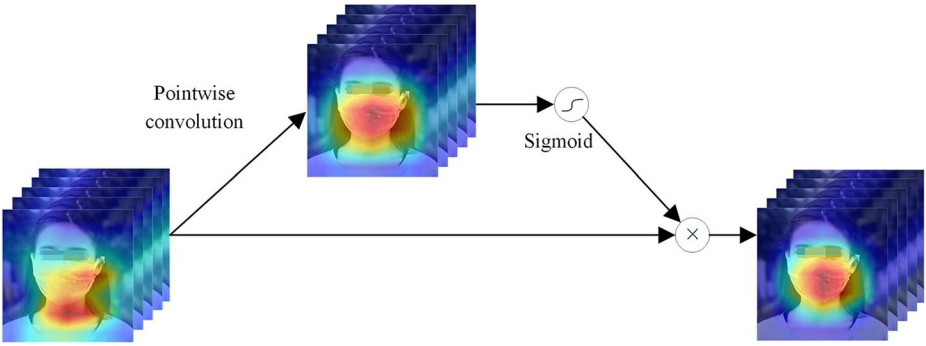

Fig. 8 Spatial attention module

whether the predicted box i and ground truth box j match in category k. l and $\hat{g}$ are the prediction box and the ground truth box, respectively. The loss regresses to offsets for the center (cx, cy) of the default bounding box and for its width (w) and height (h).

The accuracy of SSD [23] is not as high as the accuracy of the two-stage method such as FPN [21] and Cascade R-CNN [3], because SSD [23] has a problem of class imbalance in the training process. A novel loss function called focal loss is proposed to solve the class imbalance [22]. Focal loss [22] is based on cross entropy loss, making the model focus hard negatives during training by reducing the weight of easy examples. The confidence loss of SSD [23] is based on softmax loss, and we use focal loss instead of it to balance the samples. Focal loss (FL) is defined as:

$$FL(p) = -\alpha_t(1-p_t)^{\gamma}\log(p_t), p_t = \begin{cases} p & , y = 1 \\ 1-p & , y = 0 \end{cases} \tag{5}$$

where t is the sample index, p_t is the predicted probability for the class with label y = 1. α_t and γ are constant (α_t = 0.25 and γ = 2). Therefore, we get the confidence loss.

$$L_{conf} = \sum_{i\in Pos}^{N} x_{ij}^{k} FL\left(\widehat{p}_i^{\,k}\right) + \sum_{i\in Neg} FL\left(\widehat{p}_i^{\,0}\right)\widehat{p}_i^{\,k} = \frac{\exp\left(p_i^k\right)}{\sum_k \exp\left(p_i^k\right)} \tag{6}$$

3.2 SSAL

Although there are many images of mask, two problems still exist when applied on-site. First, manual labeling takes a lot of time and automatic labeling causes many errors. Second, the sample data has a high degree of similarity, so the marginal benefit generated in training is very small. A large number of similar samples cause sample imbalance, making it more difficult to detect minority objects. To solve these problems, we propose an active learning algorithm combined with semi-supervised learning called SSAL.

As shown in Fig. 9, we demonstrate the process of annotating samples by various methods. The labeling sample process of standard active learning is shown in Fig. 9a. Although active learning can select valuable samples, manual labeling takes a lot of time. In Fig. 9b, semi-supervised learning automatically labels samples to save labeling time. However, semi-supervised learning cannot distinguish the importance of samples, and there may be wrong sample annotations. As shown in Fig. 9c, the proposed algorithm uses a query strategy to select valuable samples, which are automatically marked by semi-supervised learning. Then, the experts check and modify the samples. The semi-supervised active learning algorithm is described in detail in Table 1. The proposed algorithm greatly reduces the time of expert annotation and improves the efficiency of annotation.

It is very significant to set a suitable query strategy for screening valuable samples in active learning. The query strategy we designed is to calculate the average confidence of each class and choose the smallest confidence among all classes. The query strategy is defined as:

$$y = min\left(\overline{x}_i\right) \tag{7}$$

where $\overline{x}_i$ is the average confidence of i-th class. If the y of an image is less than 0.3, the image is saved, and the corresponding class and localization are also saved for automatic labeling with semi-supervised learning. Although this query strategy retains samples with low average

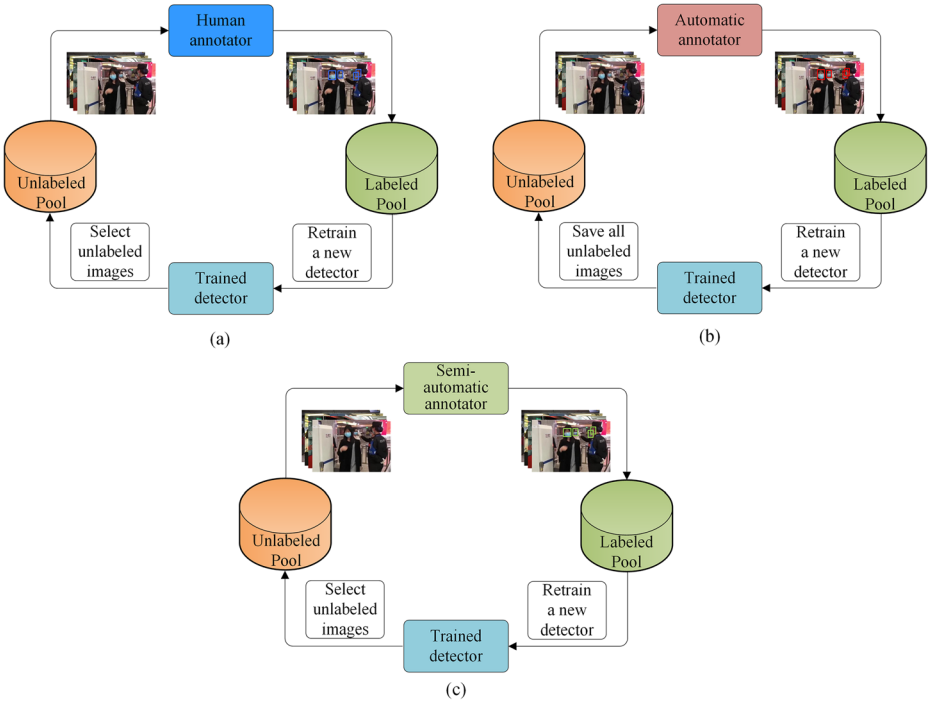

Fig. 9 Annotating samples of various methods. **a** Standard active learning; **b** Semi-supervised learning; **c** Semi-supervised active learning

Table 1 Semi-supervised Active Learning Algorithm

Algorithm：Semi-supervised Active Learning
Input: The trained detector and the unlabeled pool.
Output: The labeled pool and the retrained detector.
Step 1: Obtain the confidence and coordinates The trained detector processes the images in the unlabeled pool to obtain the confidence and coordinates of each face and mask.
Step 2: Choose the valuable images; Calculate the average confidence of the face and mask on each image separately, and choose the smallest average confidence as the ***y***, and save the images with ***y*** less than the threshold.
Step 3: Label the saved images automatically; The coordinate information obtained in the *Step 1* is used to label the saved images automatically.
Step 4: Check the labeled images manually; Check the labeled images manually, modify or delete the incorrectly labeled data, and finally get the labeled pool.
Step 5: Retrain the model; The labeled pool is used to retrain the model to obtain a new model called the retrained detector.

confidence, the selected samples also include many high-confidence objects, which ensure the diversity of the samples.

3.3 Datasets

We make a mask dataset consisting of various scenes images to train SSD-Mask. There are 25,631 images in this mask dataset, of which 22,631 images are used as the training set and 3000 images are used as the testing set. The dataset has two class: “face” and “mask”. “face” is a person who does not wear a mask, and “mask” is a person who wears a mask. A part of images from the mask dataset are shown in Fig. 10a. Our dataset has diverse and rich images including occlusion, irregular wearing, dense scenes, mask diversity, small faces and blurring. Both the occluded faces and faces that are not completely covered by masks are “face” class. The dataset contains a large number of faces, so we have not made the dataset public considering personal privacy. If any scholars need it, please contact us.

Firstly, we collect 5000 images of rail transit. Secondly, after the training of SSD-Mask is completed, we use SSD-Mask as a detector and SSAL to screen the valuable images of rail transit to make a mask dataset of rail transit. Finally, SSAL screens 1992 valuable samples as mask dataset of rail transit. Several samples in the mask dataset consisting of rail transit images are shown in Fig. 10b. Finally, we employ the mask dataset to retrain the SSD-Mask.

3.4 Training details

We implement SSD-Mask using the PyTorch and trained the network with the NVIDIA Tesla P100 GPU. Table 2 presents the parameters of SSD-Mask in training. Besides, we train Faster

(a)

(b)

Fig. 10 Mask dataset: (**a**) The mask dataset consisting of various scenes images; (**b**) The mask dataset only consisting of rail transit images

R-CNN [35], SSD [23], YOLOv3 [34] and YOLOv4 [1] on the same dataset as the comparison algorithms.

4 Results and discussion

In this section, we test the proposed algorithm and the state-of-the-art algorithms on the proposed testing set, the public mask dataset [4] and the rail transit dataset. Then, the experimental results are analyzed and discussed in detail.

4.1 Evaluation metrics

To comprehensively evaluate the algorithm results, we adopted average precision (AP), mean average precision (mAP) and frame per second (FPS) as evaluation metrics. AP is defined as

$$AP = \int_0^1 P(R)dR \tag{8}$$

where P is precision and R is recall. P and R can be formulated as

$$P = T_P/(T_P + F_P) \tag{9}$$

$$R = T_P/(T_P + F_N) \tag{10}$$

where T_P,F_P and F_N denote true positives, false positives and false negative, respectively. The mAP is mean of AP and can be calculated as

$$mAP = \sum_c AP/c \tag{11}$$

where c is the number of class.

4.2 Experimental results

4.2.1 Results of SSD-mask

To clarify the effect of SSD-Mask, we test Faster R-CNN [35], YOLOv3 [34], SSD [23], YOLOv4 [1], AIZOO [4], Baidu [42] and SSD-Mask without SSAL on the proposed testing

Table 2 Hyper-parameters of SSD-Mask in training

Parameter	Describe	Value
F	Input image size	512×512
C	Input channels	3
B	Batch size	16
I	Iteration	120,000
M	Momentum	0.9
L	Learn rate	0.001,0.0001
Ls	Learn rate step	70,000,100,000
D	Decay	0.0005
NP	Negative/positive	3

Table 3 Results on the proposed testing set

	Faster R-CNN [35]	YOLOv3 [34]	SSD [23]	YOLOv4 [1]	AIZOO [4]	Baidu [42]	SSD-Mask
mAP(%)	79.48	85.5	84.91	**86.7**	83.53	63.71	**86.34**
FPS	0.5	21	**49**	17.5	46	1.5	**42**

set including 3000 images. Note that the training code of AIZOO [4] and Baidu [42] do not open, so we have to use their trained weights. The detection confidence and intersection-over-union of all algorithms are 0.4 and 0.5, respectively. Table 3 shows the performance of SSD-Mask is better than Faster R-CNN [35], YOLOv3 [34], SSD [23], AIZOO [4] and Baidu [42].

Fig. 11 Several detection results of various algorithms on the proposed testing set. (**a**) input (**b**) Faster R-CNN [35] (**c**) YOLOv3 [34] (**d**) SSD [23] (**e**) YOLOv4 [1] (**f**) AIZOO [4] (**g**) Baidu [42] (**h**) SSD-Mask

Table 4 Results on public mask dataset

Method	Backbone	Input size	AP (%)	
			Face	Mask
Faster R-CNN [35]	ResNet-50	~600×1000	86.40	93.20
YOLOv3 [34]	Darknet-53	416×416	90.45	93.50
SSD [23]	VGGNet	512×512	90.62	90.32
YOLOv4 [1]	CSPDarkNet53	416×416	88.36	**93.60**
AIZOO [4]	ConvNet	360×360	88.81	90.04
Baidu [42]	VGGNet	128×128	54.35	76.10
SSD-Mask	VGGNet	512×512	**90.75**	91.35

Our results indicate that the proposed method is superior to them even if it did not rely on SSAL. SSD-Mask is far faster than YOLOv4 [1] with sacrificing a little bit of accuracy.

Figure 11 demonstrates several detection results of various algorithm on the proposed testing set. In the first row, Faster R-CNN [35] has three false positives and other methods predict correctly. The second row has the occlusion face, and all algorithms predict correctly except YOLOv3 [34]. In the three row, a passenger wears mask in a wrong way. Faster R-CNN [35], YOLOv3 [34], SSD [23], AIZOO [4] and Baidu [42] detect a mask. YOLOv4 [1] detects both the face and the mask. Only our method detects a face.

Currently, there are few open source mask datasets, so we choose a mask dataset proposed by [4] as the testing dataset. The mask dataset has 1838 images, including 2019 faces and 1041 masks. As shown in Table 4, the proposed method has favorable results on the public mask dataset. In the face class, the proposed method is better than all comparison algorithms. YOLOv4 [1] obtains the best results in the mask class.

4.2.2 Results of SSD-mask with SSAL

The testing dataset of rail transit has 300 images of rail transit scenes, including 1027 faces and 327 masks. The results of all algorithms on the testing dataset of rail transit are shown in Table 5. The proposed method gets the highest mAP in the testing dataset of rail transit, which demonstrates that our algorithm had a good generalization in rail transit scenarios. Moreover, the proposed method has a very fast speed, second only to SSD [23] and AIZOO [4]. Figure 12 shows several detection results of various algorithm in rail transit scenarios. Faster R-CNN

Table 5 Results on the testing dataset of rail transit

Method	Backbone	Input size	FPS	mAP(%)	AP (%)	
					Face	Mask
Faster R-CNN [35]	ResNet-50	~600×1000	0.5	54.74	61.38	48.10
YOLOv3 [34]	Darknet-53	416×416	21	71.85	81.20	62.50
SSD [23]	VGGNet	512×512	**49**	69.42	80.47	58.38
YOLOv4 [1]	CSPDarkNet53	416×416	17.5	80.57	**88.67**	72.48
AIZOO [4]	ConvNet	360×360	46	67.41	42.19	54.80
Baidu [42]	VGGNet	128×128	1.5	73.75	73.58	73.66
SSD-Mask	VGGNet	512×512	42	76.55	82.48	70.62
SSD-Mask + SSAL	VGGNet	512×512	42	**84.26**	84.85	**83.66**

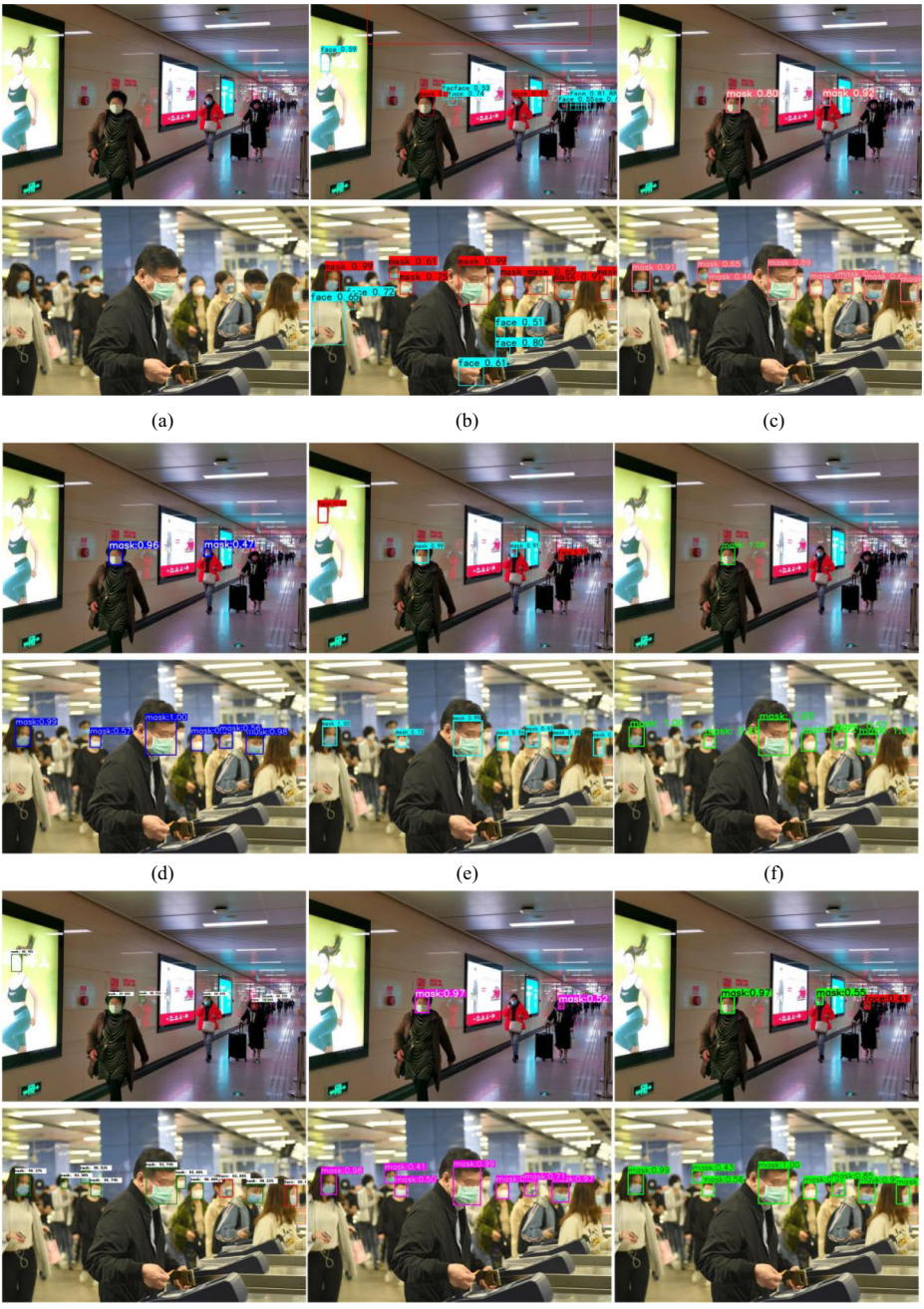

Fig. 12 Several detection results of various algorithms on rail transit scenarios. (**a**) input (**b**) Faster R-CNN [35] (**c**) YOLOv3 [34] (**d**) SSD [23] (**e**) YOLOv4 [1] (**f**) AIZOO [4] (**g**) Baidu [42] (**h**) SSD-Mask (**i**) SSD-Mask + SSAL

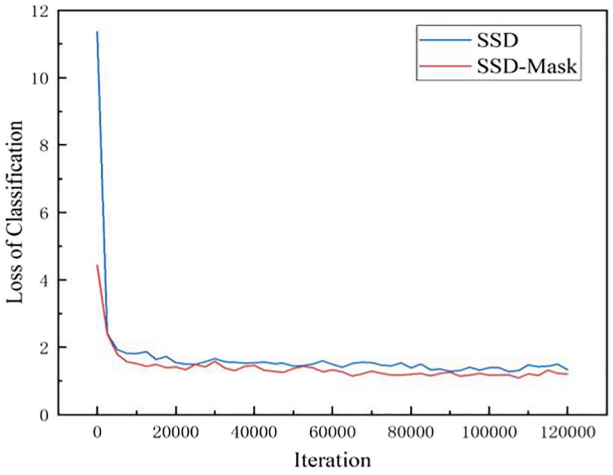

Fig. 13 The classification loss of SSD and SSD-Mask

[35] and Baidu [42] have high recall, but false positives far exceed other algorithms. The proposed method is superior to YOLOv3 [34], SSD [23], AIZOO [4], YOLOv4 [1] and SSD-Mask in both detection number and detection confidence because it is optimized for the rail transit scenario.

4.3 Discussion

4.3.1 Effect of SSD-mask

SSD-Mask has the favorable results compared with the comparison algorithms [1, 4, 34, 35, 42, 45]. As shown in Fig. 11, there are many false positives in Faster R-CNN [35], Baidu [42] and YOLOv3 [34]. It is wrong that Faster R-CNN [35] recognizes faces that are not completely covered by masks as "mask". Moreover, Faster R-CNN [35] also recognizes some backgrounds as targets. It indicates Faster R-CNN [35] recall too many negative examples. Baidu [42] has a high recall but makes the false positives increase. YOLOv3 [34] detects both occluded faces and faces that are not completely covered by masks as "mask". Unfortunately, SSD [23] and AIZOO [4] are also unable to distinguish between faces that are completely covered by masks and faces that are not completely. YOLOv4 [1] detects both the face and the mask. However, SSD-Mask can accurately detect both occluded faces and faces that are not completely covered by masks in Fig. 11. One reason is that the training dataset we made

Table 6 Ablation study of SSD-Mask

Anchor	Mish	FPN	SAM	Focal loss	mAP(%)
					84.91
√					85.50
	√				85.54
		√			85.55
			√		85.60
				√	85.52
√	√	√	√	√	**86.34**

contains occluded faces and faces that are not completely covered by masks. Another reason is that SSD-Mask has good classification performance. As shown in Fig. 13, the classification loss of SSD-Mask is smaller than that of SSD [23], which indicates that the classification performance of the former is better than that of the latter. However, all algorithms have achieved similar results on public mask dataset. This is because the public mask dataset is so simple that all algorithms can get good results.

Several ablation experiments are carried to further analyze the impact of the improved method of SSD-Mask. Table 6 proves that our proposed improvements are effective. The mAP of SSD-Mask is improved by 1.43 compared with SSD [23]. Each improvement measure is analyzed to explore how they improve the algorithm performance. Firstly, we optimize the anchor aspect ratios and activation function of backbone. The anchor aspect ratios of the SSD-Mask we set are derived from training sample statistics, so they are conducive to training the network. Mish [26] is smoother than ReLU [7, 11, 29] in Fig. 6, which may be the reason why Mish [26] is better than ReLU [7, 11, 29]. Secondly, FPN is used to fuse low-dimensional features, which is helpful for detecting small faces. Thirdly, note that SAM is the most effective of all the improvements because it allows the network to focus more on the object. Finally, we replace the softmax loss with focal loss to solve the class imbalance, which makes the network pay more attention to difficult samples during training.

4.3.2 Effect of SSAL

As shown in Table 5, SSD-Mask with SSAL shows higher performance than the state-of-the-art algorithms. These directly prove the effectiveness of SSAL. To further prove the performance of SSAL, we design a set of comparative experiments. SSAL randomly selects 1992 rail transit images for manual annotation, and these images are used to fine-tuning the SSD-Mask. Comparisons of manual annotation and SSAL on the testing dataset of rail transit is shown in Fig. 14. It is obvious that the quality of samples obtained by SSAL is better than manual annotation. The reason is that manual annotation cannot select the samples that are beneficial to the model. The labeled samples are not only repeated, but also have little effective

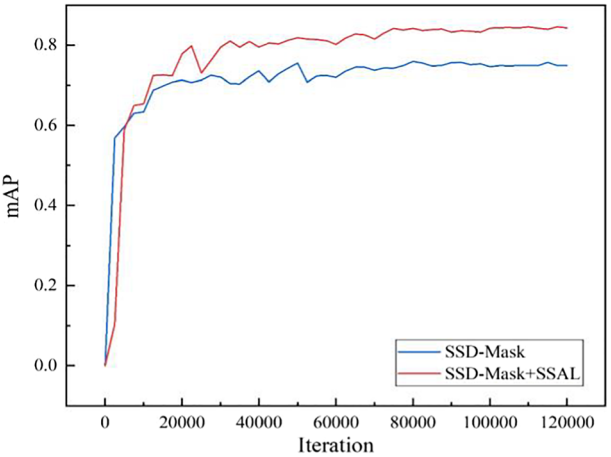

Fig. 14 The mAP on the testing dataset of rail transit

information. Moreover, SSAL greatly reduces the labeling time. It takes 4 hours to label 1992 images manually, but it takes only 0.5 hours by SSAL.

5 Conclusion

With the outbreak of the epidemic, more and more passengers wear masks at rail transit stations. This paper proposes a real-time face detection algorithm to recognize whether a face is wearing a mask, and proposes a semi-supervised active learning method to improve generalization in rail transit stations. Firstly, we propose a face detection algorithm called SSD-Mask. The performance of SSD-Mask is improved by optimizing the anchor aspect ratios, activation function and loss function, and adding feature pyramid network and spatial attention module. Secondly, this paper proposes a SSAL algorithm for screening valuable samples of rail transit and annotating these samples efficiently. Semi-supervised learning can efficiently label samples and active learning can filter valuable samples. SSAL combines these two advantages and achieves favorable results on the rail transit dataset. Extensive experiments show that the proposed algorithm has high accuracy and fast speed on rail transit dataset, which is conducive to the deployment in rail transit stations. The proposed algorithm has a wide range of applications in rail transit stations, including passenger flow statistics, epidemiological analysis, and reminders of passenger who do not wear masks. Moreover, the proposed method is not only applicable to rail transit stations, but can also be used in various public places. Finally, we respect the privacy of passengers, so our algorithm does not collect and store face information of passengers.

Authors' contributions Methodology, Writing-Original draft preparation: **Zhiwei Cao**; Conceptualization, Investigation, Funding acquisition: **Yong Qin**; Data curation, Writing-Reviewing and Editing: **Yongling Li**; Investigation: **Zhengyu Xie**; Funding acquisition: **Jianyuan Guo, Limin Jia**.

Funding This work was supported by Fundamental Research Funds for the Central Universities [2018JBZ006].

Data Availability None.

Code availability None.

Declarations None.

Conflicts of interest/competing interests None.

References

1. Bochkovskiy A, Wang CY, Liao HYM (2020) Yolov4: optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934
2. Brust CA, Käding C, Denzler J (2018) Active learning for deep object detection. arXiv preprint arXiv:1809.09875
3. Cai Z, Vasconcelos N (2018) Cascade r-cnn: Delving into high quality object detection. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 6154–6162. https://doi.org/10.1109/CVPR.2018.00644

4. Chiang D (2020) Detect faces and determine whether people are wearing mask. https://github.com/AIZOOTech/FaceMaskDetection. Accessed July 2020.
5. Desai SV, Lagandula AC, Guo W, Ninomiya S, Balasubramanian VN (2019) An adaptive supervision framework for active learning in object detection. arXiv preprint arXiv:1908.02454
6. Girshick R (2015) Fast r-cnn. In: IEEE International Conference on Computer Vision (ICCV), pp 1440–1448. https://doi.org/10.1109/ICCV.2015.169
7. Hahnloser RHR, Sarpeshkar R, Mahowald M, Douglas RJ, Seung HS (2000) Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. Nature 405(6789):947–951. https://doi.org/10.1038/35016072
8. He Y, Xu D, Wu L, Jian M, Xiang S, Pan C (2019) Lffd: a light and fast face detector for edge devices. arXiv preprint arXiv:1904.10633
9. Holub A, Perona P, Burl MC (2008) Entropy-based active learning for object recognition. In: 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp 1–8. https://doi.org/10.1109/CVPRW.2008.4563068
10. Hu J, Shen L, Albanie S, Sun G, Wu E (2019) Squeeze-and-excitation networks. IEEE Trans Pattern Anal Mach Intell:1–1. https://doi.org/10.1109/TPAMI.2019.2913372
11. Jarrett K, Kavukcuoglu K, Ranzato M, LeCun Y (2009) What is the best multi-stage architecture for object recognition? In: 2009 IEEE 12th International Conference on Computer Vision (ICCV), pp 2146–2153. https://doi.org/10.1109/ICCV.2009.5459469
12. Jeong J, Park H, Kwak N (2017) Enhancement of ssd by concatenating feature maps for object detection. arXiv preprint arXiv:1705.09587
13. Jiang M, Fan X (2020) Retinamask: A face mask detector. arXiv preprint arXiv:2005.03950
14. Joshi AJ, Porikli F, Papanikolopoulos N (2009) Multi-class active learning for image classification. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 2372–2379. https://doi.org/10.1109/CVPR.2009.5206627
15. Kao CC, Lee TY, Sen P, Liu M-Y (2018) Localization-aware active learning for object detection. In: Asian Conference on Computer Vision (ACCV), pp 506–522
16. Kovashka A, Vijayanarasimhan S, Grauman K (2011) Actively selecting annotations among objects and attributes. In: 2011 International Conference on Computer Vision (ICCV), pp 1403–1410. https://doi.org/10.1109/ICCV.2011.6126395
17. Li X, Guo Y (2013) Adaptive active learning for image classification. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 859–866. https://doi.org/10.1109/CVPR.2013.116
18. Li Z, Zhou F (2017) Fssd: feature fusion single shot multibox detector. arXiv preprint arXiv:1712.00960
19. Li J, Wang Y, Wang C, Tai Y, Qian J, Yang J, Wang C, Li J, Huang F (2019) Dsfd: Dual shot face detector. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 5055–5064. https://doi.org/10.1109/CVPR.2019.00520
20. Li X, Wang W, Hu X, Yang J (2019) Selective kernel networks. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 510–519. https://doi.org/10.1109/CVPR.2019.00060
21. Lin T, Dollár P, Girshick R, He K, Hariharan B, Belongie S (2017) Feature pyramid networks for object detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 2117–2125. https://doi.org/10.1109/CVPR.2017.106
22. Lin T, Goyal P, Girshick R, He K, Dollár P (2020) Focal loss for dense object detection. IEEE Trans Pattern Anal Mach Intell 42(2):318–327. https://doi.org/10.1109/TPAMI.2018.2858826
23. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, Berg AC (2016) Ssd: Single shot multibox detector. In: European Conference on Computer Vision (ECCV), pp 21–37. https://doi.org/10.1007/978-3-319-46448-0_2
24. Loey M, Manogaran G, Taha M, Khalifa N (2020) Fighting against COVID-19: A novel deep learning model based on YOLO-v2 with ResNet-50 for medical face mask detection. Sustain Cities Soc 65:102600
25. Long C, Hua G (2015) Multi-class multi-annotator active learning with robust gaussian process for visual recognition. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp 2839–2847. https://doi.org/10.1109/ICCV.2015.325
26. Misra D (2019) Mish: a self regularized non-monotonic neural activation function. arXiv preprint arXiv:1908.08681
27. Ml A, GMB C, MHNT D, Nemk D (2020) A hybrid deep transfer learning model with machine learning methods for face mask detection in the era of the COVID-19 pandemic. Measurement 167:108288
28. Nagrath P, Jain R, Madan A, Arora R, Kataria P, Hemanth J (2021) SSDMNV2: a real time DNN-based face mask detection system using single shot multibox detector and MobileNetV2. Sustain Cities Soc 66:102692

29. Nair V, Hinton GE (2010) Rectified linear units improve restricted boltzmann machines. In: International Conference on Machine Learning, pp. 807–814
30. Papadopoulos DP, Uijlings JRR, Keller F, Ferrari V (2016) We don't need no bounding-boxes: Training object class detectors using only human verification. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 854–863. https://doi.org/10.1109/CVPR.2016.99
31. Papadopoulos DP, Uijlings JRR, Keller F, Ferrari V (2017) Extreme clicking for efficient object annotation. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp 4940–4949. https://doi.org/10.1109/ICCV.2017.528
32. Papadopoulos DP, Uijlings JRR, Keller F, Ferrari V (2017) Training object class detectors with click supervision. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 180–189. https://doi.org/10.1109/CVPR.2017.27
33. Redmon J, Farhadi A (2017) YOLO9000: better, faster, stronger. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 7263–7271
34. Redmon J, Farhadi A (2018) Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767
35. Ren S, He K, Girshick R, Sun J (2017) Faster r-cnn: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 39(6):1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031
36. Roy S, Unmesh A, Namboodiri VP (2018) deep active learning for object detection. In: The British Machine Vision Conference (BMVC), pp 91
37. Sandler M, Howard A, Zhu M et al (2018) Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 4510–4520
38. Sethi S, Kathuria M, Kaushik T (2021) Face mask detection using deep learning: an approach to reduce risk of coronavirus spread. J Biomed Inform 5:103848
39. Settles B (2010) Active learning literature survey. University of Wisconsin, Madison 52(55–66):11
40. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556
41. Singh S, Ahuja U, Kumar M, Kumar K, Sachdeva M (2021) Face mask detection using YOLOv3 and faster R-CNN models: COVID-19 environment. Multimed Tools Appl 80(13):19753–19768
42. Tang X, Du DK, He Z, Liu J (2018) Pyramidbox: A context-assisted single shot face detector. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 797–813. https://doi.org/10.1007/978-3-030-01240-3_49
43. Uijlings J, Konyushkova K, Lampert CH, Ferrari V (2018) Learning intelligent dialogs for bounding box annotation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 9175–9184. https://doi.org/10.1109/CVPR.2018.00956
44. Vijayanarasimhan S, Grauman K (2011) Cost-sensitive active visual category learning. Int J Comput Vis 91(1):24–44. https://doi.org/10.1007/s11263-010-0372-4
45. Wang Z, Wang G, Huang B, Xiong Z, Hong Q, Wu H, Yi P, Jiang K, Wang N, Pei Y, Chen H, Miao Y, Huang Z, Liang J (2020) Masked face recognition dataset and application. arXiv preprint arXiv:2003.09093
46. Woo S, Park J, Lee J-Y, Kweon IS (2018) Cbam: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 3–19. https://doi.org/10.1007/978-3-030-01234-2_1
47. Zhang S, Wen L, Bian X, Lei Z, Li SZ (2018) Single-shot refinement neural network for object detection. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 4203–4212. https://doi.org/10.1109/CVPR.2018.00442
48. Zhang S, Chi C, Lei Z, Li SZ (2019) Refineface: refinement neural network for high performance face detection. arXiv preprint arXiv:1909.04376
49. Zhang S, Zhu R, Wang X, Shi H, Fu T, Wang S, Mei T, Li SZ (2019) Improved selective refinement network for face detection. arXiv preprint arXiv:1901.06651

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Zhiwei Cao received the bachelor's degree from China University of Mining and Technology, China, in 2017. He is currently pursuing the Ph.D. degree with Beijing Jiaotong University, China. His research interest includes traffic safety, computer vision, and image processing.

Yong Qin received the Ph.D. degree from the China Academy of Railway Sciences, Beijing, China, in 1999. He is currently with the State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University. His current research interests include the area of intelligent transportation systems, railway operation safety and reliability, rail network operation management, and traffic model.

Yongling Li received her bachelor's degree from Hebei Normal University, China, in 2019. She is currently pursuing the Master's degree in State Key Lab of Rail Traffic Control & Safety of Beijing Jiaotong University, China. Her research interests cover traffic safety, deep learning, computer vision, and image processing.

Zhengyu Xie received the Ph.D. degree from Beijing Jiaotong University, Beijing, China, in 2012. She is currently with the School of Traffic and Transportation, Beijing Jiaotong University. Her current research interests include the area of intelligent transportation systems, railway monitoring safety, and urban rail transit.

Jianyuan Guo received the Ph.D. degree from Beijing Jiaotong University, China, in 2016. She is currently the associate professor with School of Traffic and Transportation, Beijing Jiaotong University. Her research interests include Rail Transport Organization and Safety Measurement and Control.

Limin Jia received the Ph.D. degree from the China Academy of Railway Sciences, Beijing, China, in 1991.

He is currently with State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University. His current research interests include safety science and engineering, control science and engineering, transportation engineering, safety technology and engineering, and system science.