文章编号：2095-2163(2023)07-0040-06 中图分类号：TP183 文献标志码：A

# 基于多特征融合注意力的人脸口罩识别算法

安鹤男[1,2]，马 超[2]，管 聪[2]，邓武才[1]，杨佳洲[2]
（1 深圳大学 电子与信息工程学院，广东 深圳 518060；2 深圳大学 微纳光电子学研究院，广东 深圳 518060）

**摘 要**：目前，很多人出入公共场所仍须佩戴口罩。检测是否佩戴口罩变得尤为重要，而深度学习算法能够大幅度提高检测速度。本文依据 ANN 注意力机制结合特征网络改进得到 PSA(Path Strengthen Integration ANN)多尺度特征融合注意力模块，进而形成最终的 PSA-Retina 口罩识别网络，其中骨干网络基于 ResNet-50 融合空间金字塔池化；采用优化的 *GFocal Loss* 损失函数；融合 *GELUs* 激活函数重做预测器，并在口罩人脸识别数据集 RMFD 上进行对比实验，融合 PSA 模块的网络比原网络的平均精度均值 *mAP* 高 5.24%，每秒传输帧数 *FPS* 高 3.1 f/s，比 YOLOv3 网络 *mAP* 高 3.06%，*FPS* 高 0.6 f/s，实验数据表明，多特征融合注意力的 PSA-Retina 人脸口罩识别网络定位更准，准确率更高，具备在有遮挡或者非标准佩戴等情况下的检测能力，提升口罩识别效率。

**关键词**：口罩识别；ANN 注意力；PSA 模块；*GFocal Loss*

# Face mask recognition algorithm based on multi-feature fusion attention

AN Henan[1,2], MA Chao[2], GUAN Cong[2], DENG Wucai[1], YANG Jiazhou[2]
(1 College of Electronics and Information Engineering, Shenzhen University, Shenzhen Guangdong 518060, China;
2 Institute of Microscale Optoelectronic, Shenzhen University, Shenzhen Guangdong 518060, China)

**【Abstract】** At present, many people are still required to wear masks when entering and leaving public places. Detecting whether a mask is worn has become particularly important, and deep learning algorithms can greatly improve the detection speed. Based on the ANN attention mechanism and feature network improvement, this paper obtains the Path Strength Integration ANN (PSA) multi-scale feature fusion attention module, and then forms the final PSA-Retina mask recognition network. In the design, the backbone network is based on ResNet-50 fusion spatial pyramid pooling; the optimized *GFocal Loss* function is used, *GELUs* activation function is fused to redo the predictor, and a comparative experiment is conducted on the mask face recognition data set RMFD. The simulation shows that the average accuracy *mAP* of the network added to the PSA module is 5.24% higher than the original network, and the number of frames per second *FPS* is 3.1 f/s higher, *mAP* is 3.06% higher and *FPS* is 0.6 f/s higher than the results of YOLOv3 network. The experimental data demonstrates that the multi-feature fusion attention PSA-Retina face mask recognition network has more accurate positioning and higher accuracy, and has the detection ability in the case of occlusion or non-standard wearing, etc., therefore improves the efficiency of mask recognition.

**【Key words】** mask recognition; ANN attention; PSA module; *GFocal Loss*

## 0 引 言

目标检测是计算机视觉和数字图像处理的一个热门方向，不仅在监控安全、自动驾驶、工业检测、无人机场景分析等诸多领域[1-2]取得可观进展，目前也已尝试应用在口罩佩戴检测的项目及实践中[3]。该研究利用计算机视觉技术，旨在检测静止图像或视频中感兴趣的对象，对于降低人力资源成本具有重要的现实意义。具体来说，就是要识别物体属于哪个类别，更重要的是获得物体在图像中的具体位置，也可以理解为物体识别和物体定位的结合。传统的目标检测算法分别进行特征提取和分类判断，对特征选择的要求就更加严格，在面对复杂场景的时候很难得到理想效果。研究可知，时下最先进的物体检测器利用深度学习网络作为其骨干和检测网络，分别从输入图像或视频中提取特征，进行分类和定位。近几年来，随着卷积神经网络(convolution neural networks, CNN)的不断发展，目标检测算

作者简介：安鹤男(1963-)，男，研究员，主要研究方向：计算机视觉、图像处理；马 超(1998-)，男，硕士研究生，主要研究方向：目标检测；管聪(1998-)，男，硕士研究生，主要研究方向：图像处理。

收稿日期：2022-09-09

法取得了很大的突破。目前主流的算法可以分为 2 类。一类是 Two-stage 网络基于 Region Proposal 的 R-CNN 系列算法，如 R-CNN、Fast R-CNN 和 Faster R-CNN 等[4-5]，这类检测算法将检测问题划分为 2 个阶段。第一个阶段首先产生候选区域，包含目标大概的位置信息，需要先运算产生目标候选框；在第二个阶段对候选区域进行分类和位置精修。Two-stage 网络识别准确率高，漏识率低，但速度较慢，不能满足实时检测。针对这一问题，不久又研发出另一类方法，称为 One-stage，这类检测算法不需要 Region Proposal 阶段，可以通过一个 CNN 直接产生物体的类别概率和位置坐标值，已经提出的代表性算法有：YOLO、RetinaNet、SSD 等[6-7]，均可到达实时性要求。其中，RetinaNet 算法核心就是 *Focal Loss*，并在精度上超过 Two-stage 网络的精度，在速度上超过 One-stage 网络的速度，首次实现单阶段网络对双阶段网络的全面超越。近年来，对于 RetinaNet 网络的研究不断趋于深入，例如李成豪等学者[8]对小目标检测提出的 S-RetinaNet 算法，周迎峰等学者[9]提出了基于 RetinaNet 改进的海洋鱼类检测算法，由此可见 RetinaNet 算法在实际中已得到了广泛应用，但并不适用于直接检测公共场景下的口罩佩戴情况，仍然存在一些不足，亟待改进。

## 1　原理

RetinaNet 网络如图 1 所示。本次研究深入分析了极度不平衡的正负样本比例导致单阶段检测器精度低于双阶段检测器，基于上述分析，提出了一种简单、但是非常实用的交叉熵损失函数，骨干网络为 ResNet-50，特征金字塔模块接收 3 个特征图，输出 5 个特征图，通道数都是 256，步长为 8、16、32、64、128，其中大步长用于检测大物体，小步长用于检测小物体。检测头模块包括分类和位置检测两个分支，每个分支都包括 4 个卷积层，但是检测头模块的这 2 个分支之间参数不共享，分类输出通道是类别数；检测输出通道是 anchor 个数，虽然分类和回归分支权重不共享，但是 5 个输出特征图的检测头模块权重是共享的。目前存在的不足主要有以下 2 点：

（1）网络较低层级的特征层需要同时学习局部信息和高层全局信息，这种双重学习任务加大了网络训练的复杂度，进而影响检测精度。

（2）利用较低层级的检测小目标的策略完成多种尺度的目标检测，但一般情况下，深度学习神经网络中的较低层特征图能够提取充分的细节特征，但语义信息却不够丰富，反之较深的特征图包含较少的细节特征，却会造成较差的检测结果。

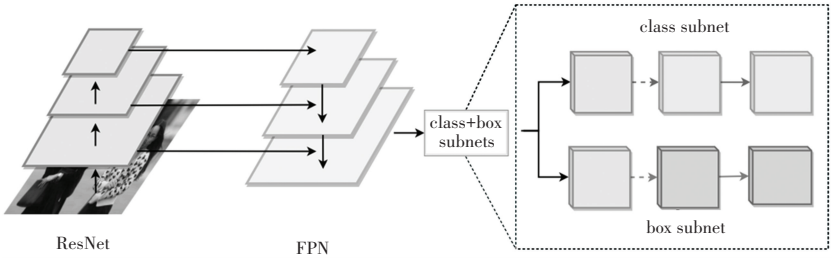



**图 1　RetinaNet 网络结构图**

**Fig. 1　RetinaNet network structure diagram**

## 2　方法与改进

本文根据以上不足以及针对特征融合能力的优化做出以下改进，基于 ResNet-50 骨干网络（backbone network）引入空间金字塔池化（Spatial Pyramid Pooling，SPP）[10]结构，利用空间金字塔池化将网络局部信息和高层信息两种学习任务加以区分，从而实现高效的目标特征学习；结合原网络多尺度特征融合想法，重新设计了 PSA（Path Aggregation Strengthen Integration ANN Attention）模块，整合链路结合多尺度特征加强融合模块，先提取丰富的局部特征，再利用自上而下和自下而上的特征融合方式将局部特征和全局特征进行融合，进一步实现特征的充分利用；采用能力更强的 *GFocal Loss* 交叉熵损失函数；根据高斯误差线性激活函数（Gaussian Error Linerar Units，GELUs）[11]，重做预测模块避免梯度爆炸，最终提出了针对人脸口罩识别的多特征融合注意力 PSA-Retina 人脸口罩识别网络，整体结构如图 2 所示。

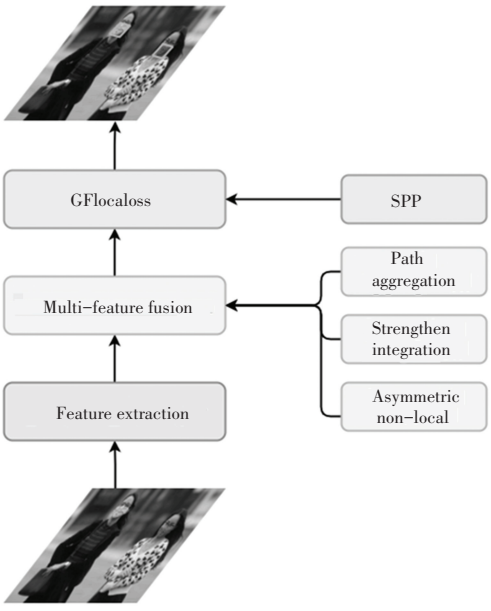



**图 2 PSA-Retina 口罩检测网络整体结构图**

**Fig. 2 The overall structure of the PSA-Retina mask detection network**

## 2.1 PSA 模块

PSA 多尺度特征融合模块主要分为 2 个子模块:整合链路模块和注意力融合模块。

整合链路模块首先要把 4 层特征的尺寸调整,$P_3$ 直接作为 $M_3$,再通过 2 倍下采样操作与 $P_4$ 相加,通过核为 3 的卷积后得到 $M_4$,$M_5$ 同理为 $M_4$ 和 $P_5$ 计算得到,$M_6$ 则为 $P_6$ 直接输出,得到 256 通道的特征图。因为底层特征分辨率较高,所以专注于细节特征的学习,顶层特征分辨率较低,总是专注于语义特征的学习。为了平衡该特性,并对特征做进一步融合,采用求和均值来计算,就是先将 4 层特征中的 $M_3$ 进行下采样,$M_5$、$M_6$ 进行上采样,保持与中间层次 $M_4$ 特征图的尺寸相一致,再进行融合处理,综上方法的数学公式见如下:

$$M = \frac{1}{L}\sum_{l=l_{\min}}^{l_{\max}} M_l \tag{1}$$

其中,$L$ 表示特征层的层数。

注意力模块将进一步处理融合后的特征图,使得特征更加有辨别力,引入 Asymmetric Non-local 注意力机制公式见如下:

$$N_i = \frac{1}{C(M)}\sum f(M_i, M_j)\, g(M_j) \tag{2}$$

其中,$N$,$M$ 特征图尺寸一致;$i$ 为输入特征图内某个元素的方位信息;$j$ 为所有可能方位信息的索引;$g$ 为信息变换函数;卷积核为 1;通过 $f$ 函数计算第 $i$ 例方位信息和其余全部方位信息的匹配性,是注意力匹配函数。用匹配计算融合后的特征图直接输出为 $N_4$,融合后的特征图再通过上采样的方法算出 $N_3$,融合后的特征图再通过下采样的方法算出 $N_5$ 和 $N_6$。共输出 4 层特征 $N_3$、$N_4$、$N_5$、$N_6$,最终与 $M$ 层特征图来计算求和,如图 3 所示,不同阶段的多尺度特征信息经过整合链路模块进行了有效的增强融合。

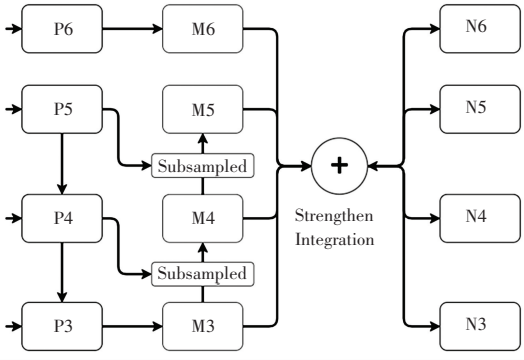



**图 3 PSA 多尺度特征融合注意力模块**

**Fig. 3 PSA multi-scale feature fusion attention module**

## 2.2 特征提取模块

骨干网络负责计算获得特征图,在 ResNet-50 网络第一个预测特征层中引入空间金字塔池化(Spatial Pyramid Pooling, SPP)结构如图 4、图 5 所示,得到表达力更强、包含多尺度目标区域信息的卷积特征图。首先,使用卷积操作将输入进来的特征处理 3 次;随后,在池化层中,对于 5、9、13 三种不同尺寸的池化核、步距为 1,分别进行最大池化下采样操作。将处理得到的特征图通过 SPP 进行拼接后,接下来将经过 3 次卷积操作,就可得到不同尺度的特征融合的输出特征图。

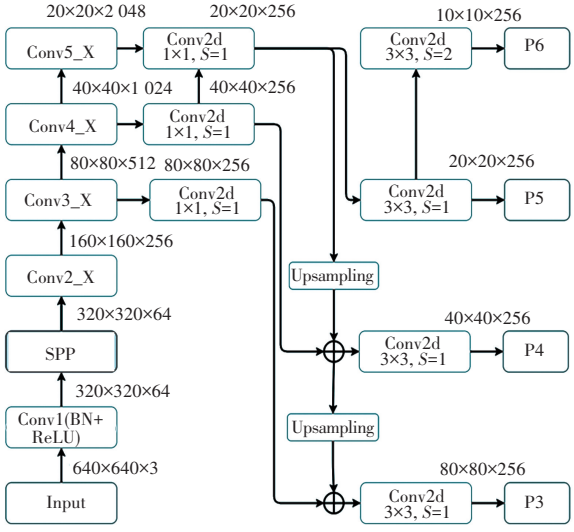



**图 4 骨干网络结构图**

**Fig. 4 Backbone network structure diagram**

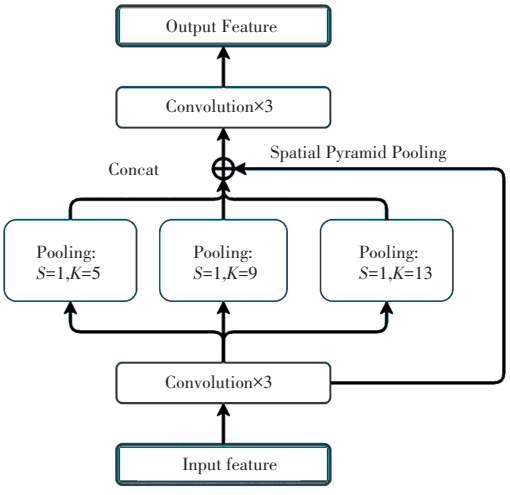



**图 5　空间金字塔池化网络结构图**

**Fig. 5　Spatial pyramid pooling network structure diagram**

### 2.3　预测器

改进预测器结构如图 6 所示。考虑通过加入正则化来提高泛化能力而避免过拟合，但是仍然存在梯度爆炸问题。为解决这一问题引入 *GELUs* 函数，其依据中心极限定理，大量独立随机变量的总体是服从近似正态分布的，现实中有很多复杂人脸口罩情况可以被建模成近似正态分布，使用类正态分布函数作为激活函数就更加合理，而且在具有相同方差的所有可能的分布中，正态分布具有最大不确定性、即熵最大。本文中，将 Class Subnet 和 Box Subnet 中的 8 个 3×3 的卷积层后的加入 *GELUs* 函数。

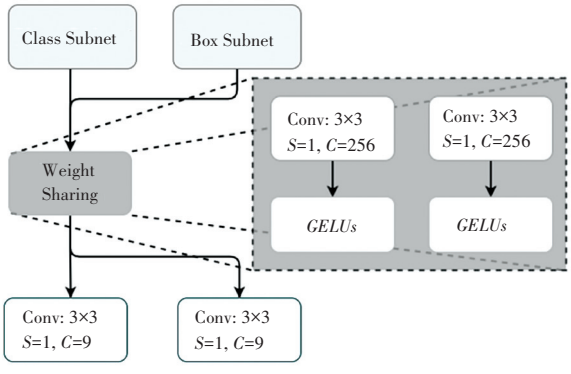



**图 6　改进预测器结构图**

**Fig. 6　Structure of the improved predictor**

### 2.4　损失函数

*GFocal Loss* 交叉熵损失函数能够有效判断真实检测框与预测检测框之间的重合度，解决了正负样本不匹配问题，进行梯度回传。为了保证训练和测试一致，同时还能够兼顾分类分数和质量预测分数都能够训练到所有的正负样本。*GFocal Loss* 将两者的表示进行联合，保留分类的向量，对应类别位置的置信度改为质量预测的分数，用离散化的方式直接回归一个任意分布来做建模框的表示，这里涉及到的数学公式为：

$$QFL = -|y-\sigma|^{\beta}((1-y)\log(1-\sigma)+y\log(\sigma)) \tag{3}$$

其中，$y$ 为标签，$\beta$ 为超参。从物理上来讲，依然还是保留分类的向量，但是对应类别位置的置信度的物理含义不再是分类的分数，而是改为质量预测的分数。由 $\delta$ 分布转为通用分布的形式：

$$y = \int_{-\infty}^{+\infty} \delta(x-y)x\mathrm{d}x \tag{4}$$

离散化后，可得：

$$y = \int_{y_0}^{y_n} P(x)x\mathrm{d}x \tag{5}$$

$$\hat{y} = \sum_{i=0}^{n} P(y_i)y_i \tag{6}$$

为了尽快拟合到真实分布，使用 *DFL*。研究推得的数学公式如下：

$$DFL = -((y_{i+1}-y)\log(S_i)+(y-y_i)\log(S_{i+1})) \tag{7}$$

其中，$y_0$ 到 $y_n$ 为积分区域；$y$ 为标签点；$S_i$ 为激活函数后的结果；$y_i$ 以及 $y_{i+1}$ 为靠近真实位置的左右邻近。在此基础上，研究推得：

$$GFL = QFL + DFL \tag{8}$$

*QFL* 和 *DFL* 的作用是正交的，两者的增益互不影响，可以统一地表示为 *GFL*。式(7) 中，$y$ 为 0 ~ 1 的质量标签；*QFL* 的全局最小解即是 $\delta = y$，实验中发现一般取 $\beta = 2$ 为最优。

## 3　实验分析

### 3.1　训练配置

训练所使用的服务器配置见表 1。本文使用 SSD、RetinaNet、YOLOv3 和 PSA-Retina 网络在同一数据集上历经相同参数的训练后进行比较，评价指标为平均精度(Average Precision，*AP*)、平均精度均值(mean Average Precision，*mAP*) 和每秒传输帧数(Frames Per Second，*FPS*)。

**表 1　服务器环境和参数**

**Tab. 1　Server configuration and environment**

| 内容 | 服务器配置 |
|---|---|
| CPU | Intel Xeon E5-2620 v4 |
| GPU | GTX 1080 Ti × 4 |
| 操作系统 | CentOS Linux release 7.61810(Core) |
| 模型使用框架 | Pytorch 1.7，python 3.7，Conda 10.1 |

### 3.2 训练数据集

武汉大学国家多媒体软件工程技术研究中心制作的 RMFD 数据集结合本次研究中经网络下载整理清洗和标注处理的真实口罩人脸识别数据集，该数据集包含 6 000 张口罩人脸和 91 000 张不戴口罩人脸，挑选其中 5 000 张戴口罩，5 000 张不戴口罩、总共 10 000 张图片，并将该数据集的 60%用作训练集，40%用作测试集。基于平均检测精度、平均精度均值及运行帧率评价指标，将提出的 PSA-Retina 网络进行实验和评估，设置 *IoU* 为 0.4，所有模型训练批尺寸设置为 8，初始学习率设置为 0.000 1，训练 200 个周期。

### 3.3 实验结果与分析

不同网络实验结果见表 2。基于多特征融合的 PSA-Retina 网络对戴口罩的 *AP* 值达到 87.21%，对未佩戴口罩的 *AP* 值达到 83.05%，检测器的 *mAP* 值达到 85.13%，*FPS* 值达到 33.7 f/s。本文提出的网络要比 SSD、RetinaNet、YOLOv3 网络的检测精度均高出 3%以上，且检测速度也表现最优，说明本网络结构更适用于口罩的识别检测。

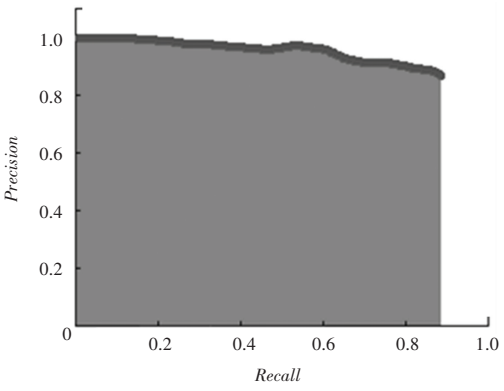
**表 2 不同网络实验结果**

**Tab. 2 Experimental results of different network**

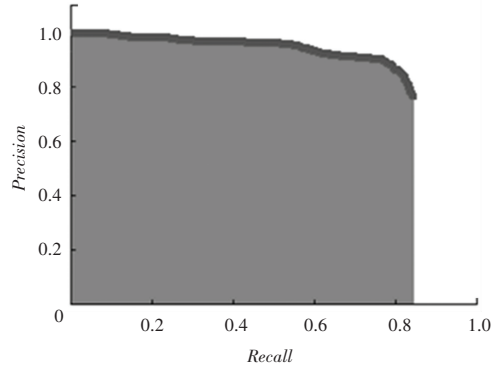
| 方法 | *AP*/ % | | *mAP*/ % | *FPS*/ (f · s$^{-1}$) |
|---|---|---|---|---|
| | mask | unmask | | |
| SSD | 85.30 | 73.02 | 79.16 | 33.3 |
| RetinaNet | 80.52 | 79.26 | 79.89 | 30.6 |
| YOLOv3 | 87.54 | 76.60 | 82.07 | 33.1 |
| PSA-Retina | 87.21 | 83.05 | **85.13** | **33.7** |

将训练后的网络在测试集上进行测试，获得了召回率-精确度（*P* - *R*）曲线如图 7 所示。图 7 中，曲线下围成的面积即为平均检测精度 *AP*。

为更加直观地感受 PSA-Retina 网络对口罩识别的有效性，图 8 展示了 SSD 网络、RetinaNet 原网络、YOLOv3 网络、改进的 PSA-Retina 网络在人脸口罩数据集的检测效果对比结果，其中置信度阈值设置为 0.4，非极大值抑制 *NMS* 阈值设为 0.45。由对比结果可以看出，SSD 网络对小目标的检测效果并不好；RetinaNet 网络相比 SSD 略有提升，但是容易漏检，有些很明显的目标反而没有被检测到；YOLOv3 网络效果较好，但仍然有漏框出现；本文提出的算法对小目标和遮挡都表现出良好的效果，绝大部分漏检、错检情况都被修复，更加适合实际应用中高检测精度的需求。




（a）Class：86.21% = mask *AP*




（b）Class：82.05% = unmask *AP*

**图 7 检测召回率-精度曲线图**

**Fig. 7 Detection *Recall-Precision* plot**

（a）SSD （b）RetinaNet

（c）YOLOv3 （d）PSA-Retina

**图 8 不同网络检测效果对比图**

**Fig. 8 Comparison chart of different network detection effects**

## 4 结束语

本文根据ResNet-50和SPP空间金字塔结构

（下转第 52 页）