文章编号:1674-2869(2024)04-0424-07

面向边缘设备的轻量级OpenPose姿态检测模型研究

黄瑜豪[1,2],曾祥进*[1,2],冯 崧[1,2]
1. 武汉工程大学计算机科学与工程学院,湖北 武汉 430205;
2. 湖北三峡实验室,湖北 宜昌 443007

摘 要:为了在低算力的边缘设备上使人体行为识别网络兼顾实时性和识别效果,提出了一种改进的轻量级OpenPose姿态检测模型。使用移动网络替换原主干特征提取网络,在特征提取网络的浅层使用倒置残差结构,减少浅层网络的运算量,在网络深层引入卷积块注意力模块,调整深层特征信息的权重,并将浅层网络特征与深层网络特征融合后送入卷积网络进行骨骼关键点的拼接,有效融合浅层和深层的特征信息。在COCO数据集上的验证结果表明:改进模型与原模型相比,正确关键点百分比提升了2.8%,平均精度提升了2.0%。使用改进后的模型作为预训练模型在行为数据集上标记骨骼关键点用作分类训练,将完成分类训练的模型部署在边缘设备上,在边缘设备运行速度略微降低的情况下,通过改进后的模型进行人体行为识别的准确率达到96.4%,有效实现在边缘设备上的姿态检测和人体行为识别。

关键词:人工智能;OpenPose;移动网络;机器视觉
中图分类号:TP391.4 **文献标识码**:A **DOI**:10.19843/j.cnki.CN42-1779/TQ.202311008

A lightweight OpenPose posture detection model for edge devices

HUANG Yuhao[1,2], *ZENG Xiangjin*[*1,2], *FENG Song*[1,2]
1. School of Computer Science and Engineering, Wuhan Institute of Technology, Wuhan 430205, China;
2. Hubei Three Gorges Laboratory, Yichang 443007, China

Abstract: In order to achieve real-time human behavior recognition on edge devices with low computational power while keeping a balance between real-time performance and recognition effectiveness, we proposed an improved lightweight OpenPose posture detection model. The model replaces the original backbone feature extraction network with mobile networks, uses inverted residual structures in the shallow layers of the feature extraction network to reduce computational complexity, introduces convolutional block attention module in the deep layers to adjust the weights of deep feature information, and fuses shallow and deep feature information after merging them before feeding them into the convolutional network for the concatenation of skeleton keypoints. This effectively integrates both shallow and deep feature information. Validation results on the COCO dataset show that compared to the original model, the improved model achieves a 2.8% increase in correct keypoint percentage and a 2.0% increase in average precision. Using the improved model as a pre-trained model, skeletal keypoints are labeled on a behavioral dataset for classification training. When deploying the trained model on edge devices, even with a slight decrease in operating speed on edge devices, the accuracy of human behavior recognition reached 96.4%, effectively realizing posture detection and human behavior recognition on edge devices.

Keywords: artificial intelligence; OpenPose; mobile networks; computer vision

收稿日期:2023-11-15
基金项目:国家自然科学基金(61502355);湖北省湖北三峡实验室创新基金(SC215001)
作者简介:黄瑜豪,硕士研究生。Email:hyh1184@163.com
***通信作者**:曾祥进,博士,副教授。Email:xjzeng21@163.com
引文格式:黄瑜豪,曾祥进,冯崧. 面向边缘设备的轻量级OpenPose姿态检测模型研究[J]. 武汉工程大学学报,2024,46(4):424-430.

随着计算机硬件能力的提升，人工智能在视觉领域的发展尤为迅速[1]，其中基于机器视觉的目标姿态检测方法正逐渐受到人们的关注，并广泛应用在各种领域，例如检测操作工人的异常行为[2]、独居老人的摔倒行为[3]、公共环境的暴力行为[4]等。现有基于机器视觉的人体姿态检测算法包括传统的图像算法和基于深度学习的姿态检测算法。传统的图像算法通过图像分割、特征变换等方式提取目标特征做检测，这种方法识别准确率高、速度快，但是针对不同的特征需要重新适配，泛用性较差。基于深度学习的姿态检测算法中应用最广泛的是OpenPose[5]模型结构，例如Ramesh等[6]使用OpenPose模型对跑步行为进行动作捕捉，但该算法模型的部署严重依赖高算力的设备，对小开发团队或者个人开发者极其不友好；李一凡等[7]使用轻量级残差网络（resident networks，ResNet）替换原网络，有效提升了OpenPose模型的运行速度，降低了模型参数量，但是分类效果不佳；王非等[8]使用轻量化的移动网络（mobile networks，MobileNet）替换OpenPose模型中的特征提取网络，进一步降低了模型参数量，但是检测效果不够理想。

为了在低算力的边缘设备上部署基于深度学习的姿态检测网络，综合考虑网络识别效果和实时性的需求，实现基于姿态检测的人体行为识别，本文研究并设计了一种轻量级OpenPose模型。在特征提取层使用改进的MobileNet替换原本的主干特征网络，通过在浅层使用ResNet进一步降低模型训练的参数量，同时为了保证模型的检测效果，在网络深层引入卷积块注意力模块（convolutional block attention module，CBAM）[9]，在姿态估计层重新设计网络结构。改进后的轻量级OpenPose模型能够在边缘设备上实时部署，并准确识别目标场景下的人体行为。

1　行为识别平台设计

平台使用的嵌入式设备为英伟达旗下的Jetson TX2设备。Jetson TX2设备包含256个并行计算架构（compute unified device architecture，CUDA），可以对神经网络模型进行推理。平台与网络摄像头需接入同一局域网内，平台通过交换机读取摄像头的实时视频流，并将其逐帧输入模型，运行在Jetson TX2设备的CUDA加速核心上。在完成神经网络的优化、部署工作后，将结果输出至显示界面，该系统的工作原理如图1所示。

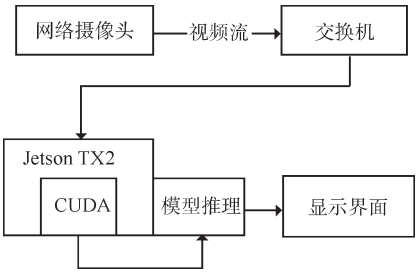

图1　系统工作原理

Fig. 1　Operating principle

2　姿态检测算法

2.1　OpenPose姿态检测模型

OpenPose模型是2016年由卡耐基梅隆大学开源的一项人体姿态检测模型，其在多人图像中使用自底向上的方法对人体姿态进行识别。首先确定骨骼关键点的位置信息，并将其标记为关节热度图，然后通过计算每个关节热度图对应的局部亲和场（part affinity fields，PAF）将各个关键点连接构成向量。传统的OpenPose模型使用视觉几何组网络（visual geometry group network，VGGNet）[10]作为输入层，VGGNet包含16个卷积层以及3个全连接层。原始图像输入网络后在前10层生成特征图$\boldsymbol{F}$，并将特征图$\boldsymbol{F}$作为后续神经网络的输入。后续网络中每个阶段的姿态检测层会分出2个分支层，在每个阶段，分支层1用来预测关节热度图S_t，分支层2用来预测局部亲和场L_t，其中t表示网络推理过程中的阶段数，具体如式（1）所示：

$$\begin{aligned} S_t &= \rho_t(\boldsymbol{F}, S_{t-1}, L_{t-1}), \forall t \geqslant 2 \\ L_t &= \varphi_t(\boldsymbol{F}, S_{t-1}, L_{t-1}), \forall t \geqslant 2 \end{aligned} \tag{1}$$

式（1）中：ρ_t和φ_t分别表示两个分支层在第t阶段用于推理的卷积网络，每个阶段的输入为$\boldsymbol{F}$和前一阶段预测的结果，使用本阶段的卷积网络分别预测出本阶段的关节热度图和PAF。

VGGNet在整个网络中全部使用3×3大小的卷积核堆叠神经网络。使用VGGNet的OpenPose模型网络结构如图2所示，其中，h，w分别表示输入图像的高和宽；h'，w'分别表示输出图像的高和宽；C表示卷积操作；P表示池化操作。

2.2　改进的轻量级OpenPose模型

2.2.1　深度可分离卷积　由于Jetson TX2设备的推理性能不高，基于VGGNet的OpenPose模型在该设备上无法正常运行，本文将原本的主干特征网络全部替换为MobileNet[11]，后者使用深度可分离卷积的思想，将标准卷积拆分成1个普通卷积层

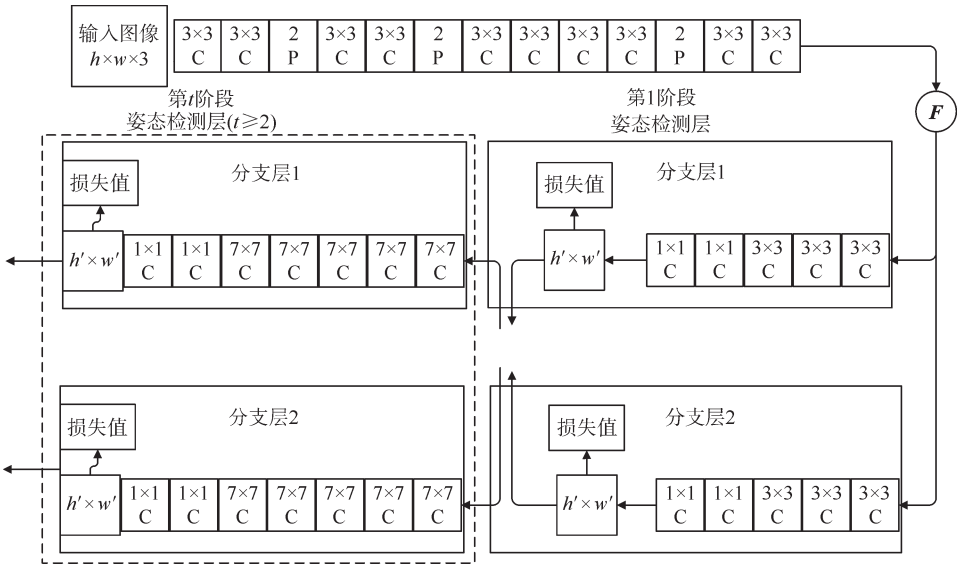

图2 OpenPose模型的网络结构

Fig. 2 Network structure of OpenPose model

和1个1×1的点阵。假设输入层特征矩阵为$D_F \times D_F \times M$,输出层特征矩阵为$D_F \times D_F \times N$,卷积核为$K \times K$,其中D_F表示拆分成方形的输入特征图像和输出特征图像的边长,M和N分别表示输入特征和输出特征的通道数,K表示卷积核尺寸。标准卷积的计算量S_1和深度可分离卷积的计算量S_2如式(2-3)所示:

$$S_1 = D_F \times D_F \times M \times N \times D_F \times D_F \quad (2)$$

$$S_2 = K \times K \times M \times D_F \times D_F + M \times N \times D_F \times D_F \quad (3)$$

对比S_1和S_2的参数量可知,深度可分离卷积的计算量仅为标准卷积计算量的$1/N+1/K^2$,N通常是一个很大的值,因此当K=3时,深度可分离卷积的计算开销仅为标准卷积的1/9。

2.2.2 特征融合 在网络浅层使用倒置残差(inverted residuals,IR)结构,该结构具有瓶颈设计,如图3所示,其中c表示通道数,a表示扩张倍率,图3中公式表示传递过程中$\boldsymbol{F}$通道数的变化。特征图$\boldsymbol{F}$作为输入首先通过一次卷积将通道扩张a倍进行升维,实现特征从低维到高维的表征,然后使用激活函数分配各通道信息的权值,抑制非特征通道,再经过一次卷积将通道缩减为$1/a$进行降维,将高维通道信息映射回低维通道[12],残差块经过"升维-卷积-降维"的步骤,使输出特征图$\boldsymbol{F}'$随着卷积层的加深,实现层间跨通道的信息传递,有效保留通道特征[13-14]。

深度可分离卷积的引入虽然能极大减少模型运算量,但同时也很容易导致卷积过程中特征信息的丢失,为了保证网络深层中姿态检测层PAF

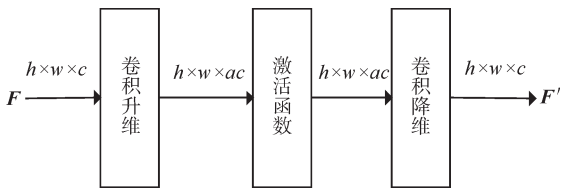

图3 IR结构

Fig. 3 IR structure

的拼接效果,在网络的深层添加CBAM,CBAM结构如图4所示。

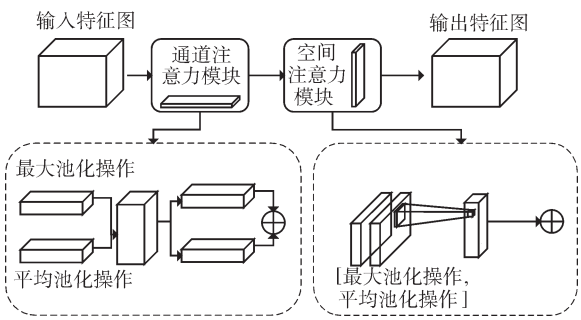

图4 CBAM结构

Fig. 4 CBAM structure

输入特征图先后经过通道注意力模块和空间注意力模块,通道注意力模块通过并行的最大池化操作和平均池化操作生成通道注意图,调整$\boldsymbol{F}$通道内的特征信息;空间注意力模块通过先后进行的最大池化操作和平均池化操作生成空间注意力图,调整$\boldsymbol{F}$坐标间的特征信息。通过为通道注意力模块和空间注意力模块分配不同的权重,对各通道和不同坐标下特征图中特征信息的表征能力进行调整,有效提高网络对特征图中特征的关

注度。

CBAM可以加强通道和空间特征信息的表征[8]，其计算量仅与通道数和卷积核大小相关，不会带来过大的计算开销，非常适合部署在边缘设备的神经网络中。

2.2.3　*姿态估计层的改进*　原本的姿态检测层在PAF的拼接过程中使用多个7×7的卷积核，本文使用3个3×3的卷积核等效替代1个7×7的卷积核，假设输入尺寸相同的图片，输入尺寸为$M\times M$，卷积核的大小为$K\times K$，填充为P，步长为S，则输出尺寸M'如式(4)所示：

$$M' = (M + 2 \times P - K)/S + 1 \qquad (4)$$

从式(4)可以看出，在相同的填充和步长下，使用3个3×3的卷积核等效替代1个7×7的卷积核时感受野不变，同时减少卷积带来的计算量。

改进后的轻量级OpenPose模型使用深度可分离卷积MobileNet替换VGGNet；在特征提取的浅层使用IR结构，深层添加CBAM，将浅层特征降采样后与深层特征融合再送入姿态检测层中的各个分支层；分支层中PAF拼接网络使用3个3×3卷积核代替7×7卷积核减少计算开销。重新设计的网络结构如图5所示，其中Cd表示深度可分离卷积操作。

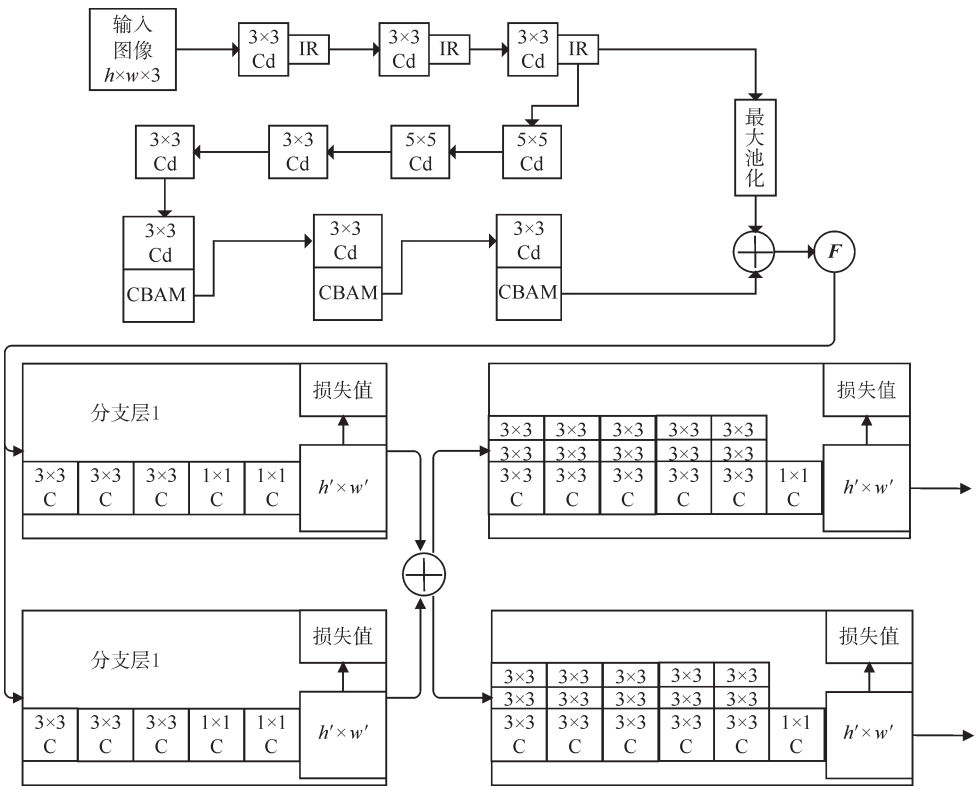

图5　改进的轻量级OpenPose网络结构

Fig. 5　Improved lightweight OpenPose network structure

3　结果及讨论

3.1　数据集

使用COCO数据集验证改进网络效果，在有人体的图片中定义了18个骨骼关键点。网络训练过程中将含有人体的4万张图片分别设置为训练集、测试集和验证集，比例为8∶1∶1。

在边缘设备上验证基于姿态检测的人体行为识别效果时，使用自行录制的行为数据集，通过骨骼关键点的空间位置差异描述不同的行为，数据集共包括8类行为，分别是拳击、鼓掌、挥手、坐、跑、走、站立、弯腰。从视频中一共提取12 000张图片，平均每类行为有1 500张图片，以在COCO数据集上训练完成的模型为预训练模型对每张图片标注骨骼关键点制作成数据集，并按照8∶2的比例制作训练集和验证集。

3.2　训练结果分析

改进后的轻量级OpenPose模型在Pytorch2.1平台上训练，平台的硬件配置包括英特尔酷睿i5-13600KF，32 GB内存和型号为英伟达RTX4080的GPU，Python版本为3.9.16。在COCO数据集上的训练效果以正确关键点百分比(percentage of correct

keypoints，PCK）、平均精度（average precision，AP）和均方误差损失（mean squared error，MSE）作为评估模型性能的指标。PCK指的是当检测的骨骼关键点位置与标注的人体关键点位置之间的距离低于某个阈值时，认为该点被正确检测到，本文中阈值设定为5像素值；AP是单类别检测精度，其计算公式如式（5）所示：

$$P_{\mathrm{A}}=\int_0^1 \frac{T_{\mathrm{P}}}{T_{\mathrm{P}}+F_{\mathrm{P}}}\left(\frac{T_{\mathrm{P}}}{T_{\mathrm{P}}+F_{\mathrm{N}}}\right)\mathrm{d}\left(\frac{T_{\mathrm{P}}}{T_{\mathrm{P}}+F_{\mathrm{N}}}\right) \tag{5}$$

其中，T_{P}表示单样本中PCK值为1的数量，即检测到的关键点数量与标签中关键点数量一致，F_{P}表示单样本中PCK值在1.0以下的数量，即检测到的关键点数量小于标签中关键点数量，F_{N}表示单样本中PCK值在1.0以上的数量，即检测到的关键点数量大于标签中关键点数量。

MSE分为局部亲和场损失值P_{Loss}和热度图损失值H_{Loss}两部分，计算方法如式（6）所示：

$$L_{\mathrm{MSE}}=\sum_{t=1}^{T}(P_{\mathrm{Loss}}+H_{\mathrm{Loss}}) \tag{6}$$

其中，T表示总时间周期，P_{Loss}表示在第t阶段姿态检测层中预测PAF产生的误差，计算公式如式（7）所示：

$$P_{\mathrm{Loss}}=\boldsymbol{W}\cdot||S_t-S_t^*||^2 \tag{7}$$

H_{Loss}表示在第t阶段姿态检测层中预测关节热度图产生的误差，计算公式如式（8）所示：

$$H_{\mathrm{Loss}}=\boldsymbol{W}\cdot||L_t-L_t^*||^2 \tag{8}$$

式（7-8）中：$\boldsymbol{W}$表示二值化矩阵，当目标中关键点的标签存在时，$\boldsymbol{W}$值为1，否则为0；S_t和S_t^*分别表示第t阶段姿态检测层中关节热度图预测的和实际的关键点坐标值，L_t和L_t^*分别表示第t阶段姿态检测层中PAF预测的和实际的关键点坐标值。

MSE表示模型中每个分支层的损失值之和，该值越低，表示预测结果与目标结果越接近，模型的预测效果越好。

为了证明改进算法的优越性，在同样的数据集上进行不同主干特征网络下的对比实验，在部分图片上的检测结果如图6所示，对比结果如表1所示。从表1中可以看出本文方法对比原方法PCK提升了2.8%，AP提升了2.0%，MSE值减小了12.4%，有效提高了骨骼关键点的检测效果，降低了预测结果与目标结果的损失；从图6中可以看出本文方法在多人场景下的姿态估计效果更优。

表1 各模型性能评估指标对比结果

Tab. 1 Comparison results of performance evaluation indicators for different models

主干特征网络	PCK / %	AP / %	MSE
MobileNetV1	76.8	53.7	146.4
MobileNetV2	77.1	54.4	138.5
MobileNetV1+IR	75.9	54.0	135.7
MobileNetV1+CBAM	77.9	55.2	133.7
本文方法	79.6	55.7	130.2

3.3 模型移植结果分析

以在COCO数据集上训练过的改进模型为预训练模型，对行为数据集标记骨骼关键点做分类训练，将训练完成后的模型移植至Jetson TX2设备上，该设备在行为数据集上推理结果的混淆矩阵如表2所示，矩阵对角线上的数值表示该行为识别的准确率，计算方法如下：只输出识别正确行为的数

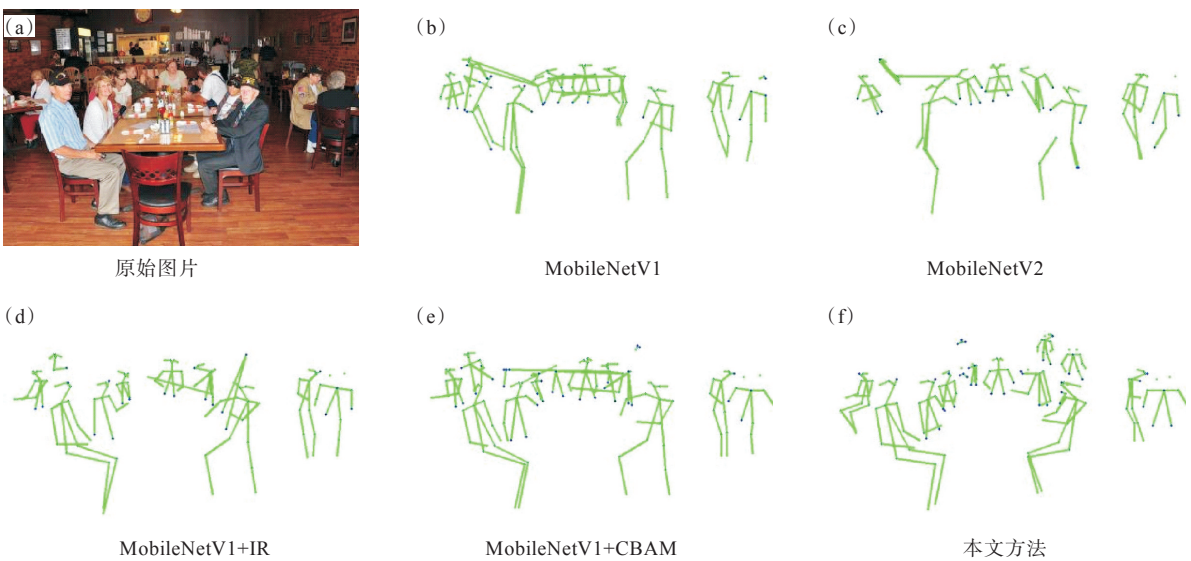

图6 不同主干特征网络下检测结果

Fig. 6 Results of different backbone feature networks

量占总数量的比例，不统计漏检的情况。从表2中可以看出模型对8类行为的综合识别率达到96.4%。

表2　行为识别混淆矩阵

Tab. 2　Behavior recognition confusion matrix

行为种类	识别率/%							
	拳击	鼓掌	挥手	坐	跑	走	站立	弯腰
拳击	99.6	0	0.4	0	0	0	0	0
鼓掌	0	93.4	3.4	0	1.6	0.4	1.2	0
挥手	0.4	1.8	95.6	0	0.2	0.2	1.8	0
坐	0	0	0	96.4	0	0.2	0	3.4
跑	0	1.2	1.6	0	95.4	1.4	0.4	0
走	0	0	0	0	3.2	95.6	1.2	0
站立	0	0	1.0	0	0.8	1.6	96.6	0
弯腰	0	0.2	0	0	0	1.2	0	98.6

将本文的改进方法与包含相同行为类别的相关行为识别研究方法进行对比，结果如表3所示。从表3中可以看出本文方法对比相关研究方法在站立和走路行为的识别准确率上略微不足，但是对坐和弯腰行为的识别效果明显提升（“/”表示对应方法中的行为识别不包括该类行为）。

本文的改进方法与原始模型及相关行为识别研究方法的平均识别效果如表4所示，可以验证本文改进方法在平均识别效果上的优越性。

表3　相关行为识别研究结果对比

Tab. 3　Comparison results of research on related behavior recognition

方法	识别率/%			
	站立	坐	走	弯腰
参考文献[6]	87.7	83.5	79.3	/
参考文献[15]	100.0	/	97.3	83.3
参考文献[16]	100.0	89.0	95.3	/
本文方法	96.6	96.4	95.6	98.6

表4　相关行为识别研究平均识别率对比

Tab. 4　Comparison of average recognition rates in related behavior recognition research

方法	行为数量	平均识别率/%
MobileNetV1	8	89.25
参考文献[6]	5	82.56
参考文献[15]	11	96.04
参考文献[16]	9	95.05
本文方法	8	96.40

部分行为的识别效果如图7所示，可以看出在简单场景下无干扰物遮挡时，行为识别模型的推理基本可以实现对骨骼关键点的检测，并且较准确地识别出人体的行为，平均速度可以达到8.8帧/s。



图7　部分行为识别效果图

Fig. 7　Partial behavior recognition rendering

4　结　论

本文面向边缘设备提出了一种改进的轻量级OpenPose姿态检测模型，在边缘设备上兼顾实时性和识别效果。在特征提取层使用轻量级的MobileNet替换VGGNet，降低推理时的参数量；在

网络的浅层使用IR结构，深层添加CBAM，融合浅层和深层的特征信息；在姿态估计层使用3个3×3卷积代替1个7×7卷积，保证感受野不变的同时降低运算量。在COCO数据集上的验证结果表明，改进方法与原方法相比，PCK提升了2.8%，AP提升了2.0%，MSE值减小了12.4%。

将该模型作为预训练模型在行为数据集上标记骨骼关键点进行分类训练，使用完成分类训练的模型部署在边缘设备上进行推理测试，可以较准确地分辨出数据集中的8类行为，平均识别率达到96.4%，平均速度达到8.8 帧/s，能够在提升识别效果的同时满足模型推理的实时性需求，但是该模型在复杂场景下的效果并未验证，未来可以作进一步的研究和优化。

参考文献

[1] 李清格，杨小冈，卢瑞涛，等. 计算机视觉中的Transformer发展综述[J]. 小型微型计算机系统，2023，44(4)：850-861.

[2] 魏麟，谭任翔，何峻毅，等. 基于改进YOLOv5的飞行员异常行为识别方法[J]. 航空计算技术，2023，53(6)：20-24.

[3] 崔悦，张德育，王国杰. 基于F-Faster-RCNN算法的摔倒检测研究[J]. 沈阳理工大学学报，2023，42(1)：1-6.

[4] 李冠，庞玉琳，田坤. 基于YOLO和ConvLSTM混合神经网络的暴力视频检测[J]. 计算机应用与软件，2023，40(11)：233-240.

[5] CAO Z, SIMON T, WEI S E, et al. Realtime multi-person 2D pose estimation using part affinity fields [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Piscataway, NJ: IEEE, 2017: 7291-7299.

[6] RAMESH S H, LEMAIRE E D, TU A, et al. Automated implementation of the edinburgh visual gait score (EVGS) using OpenPose and handheld smartphone video[J]. Sensors, 2023, 23(10): 4839.

[7] 李一凡，袁龙健，王瑞. 基于OpenPose改进的轻量化人体动作识别模型[J]. 电子测量技术，2022，45(1)：89-95.

[8] 王非，刘军. 基于MobileNet和NAM注意力机制的轻量级OpenPose网络[J]. 通信与信息技术，2023(2)：8-12.

[9] 陈江川，吴云韬，孔权. 基于CBAM-Res2Net的人群计数算法[J]. 武汉工程大学学报，2022，44(6)：664-669.

[10] PFISTER T, SIMONYAN K, CHARLES J, et al. Deep convolutional neural networks for efficient pose estimation in gesture videos [C]//Computer Vision-ACCV 2014. Berlin: Springer, 2015: 538-552.

[11] HOWARD A G, ZHU M L, CHEN B, et al. MobileNets: efficint convolutional neural networks for mobile vision applications [OL]. (2017-04-17) [2023-11-15]. https://doi.org/10.48550/arXiv.1704.04861.

[12] 刘树东，刘业辉，孙叶美，等. 基于倒置残差注意力的无人机航拍图像小目标检测[J]. 北京航空航天大学学报，2023，49(3)：514-524.

[13] 刘群坡，盛月琴，高如新，等. 基于关键帧和注意力残差网络的手语识别[J]. 计算机工程，2023，49(12)：224-230，242.

[14] ALHASSAN A M, ZAINON W M N W. Brain tumor classification in magnetic resonance image using hard swish-based RELU activation function-convolutional neural network [J]. Neural Computing and Applications, 2021, 33(15): 9075-9087.

[15] 伏娜娜，刘大铭，张恒博，等. 面向嵌入式系统的人体行为识别[J]. 激光与光电子学进展，2022，59(22)：2215001.

[16] 贾小云，王二虎，吴敬一. Android平台下的实时人体行为识别[J]. 计算机工程与应用，2018，54(24)：164-167，175.

本文编辑：霍淑利