DOI:10.3969/j.issn.1005-202X.2025.01.011

医学信号处理与医学仪器

面向家庭场景的轻量级婴幼儿姿态估计方法

万金亮[1],熊启亮[1],刘苑[2],莫杰义[1],谌颖[1]
1.南昌航空大学仪器科学与光电工程学院,江西 南昌 330063;2.重庆医科大学附属儿童医院康复科,重庆 400044

【摘要】如何有效降低婴幼儿姿态估计网络模型的大小是制约婴幼儿姿态估计技术“家用化”的关键问题,提出一种面向家庭场景的婴幼儿轻量级姿态估计方法,该方法利用轻量级网络MobileNetV3作为编码主干,并在解码部分引入基于PixelShuffle上采样模块以降低模型参数量。同时,本文通过引入坐标注意力机制(CA)的方式,能够更好地捕获位置信息和通道特征信息,突出图像中小目标和遮挡人体关键点的特征信息。最后,进一步修改并行交叉连接卷积部分以增强特征信息提取能力。在人体姿态估计通用数据集COCO以及面向婴幼儿姿态估计的专用数据集SyRIP上分别进行方法性能的验证,实验结果表明,在计算量(GFLOPs)只有0.96的情况下,在COCO和SyRIP数据集中模型平均精度分别达到73.5%、91.0%,证明本文提出的方法在显著降低模型参数和计算量的同时,不会损失姿态估计模型的准确性。本文提出的轻量化估计模型有望部署于智能终端等家用设备上,从而实现家庭场景下婴幼儿异常姿态的智能评估。

【关键词】婴幼儿姿态估计;轻量化;运动姿态;MobileNetV3
【中图分类号】R318 【文献标志码】A 【文章编号】1005-202X(2025)01-0072-10

Lightweight infant pose estimation in home scenarios

WAN Jinliang[1], XIONG Qiliang[1], LIU Yuan[2], MO Jieyi[1], CHEN Ying[1]
1. School of Instrument Science and Optoelectronic Engineering, Nanchang Hangkong University, Nanchang 330063, China; 2. Department of Rehabilitation, Children's Hospital of Chongqing Medical University, Chongqing 400044, China

Abstract: How to effectively reduce the size of infant pose estimation network models is a key issue restricting the "home-use" of infant pose estimation technology. Therefore, a lightweight method for infant pose estimation in home scenarios is proposed. The method takes the lightweight network MobileNetV3 as the encoding backbone and utilizesa PixelShuffle up-sampling module in the decoder for reducing the quantity of model parameters. Meanwhile, coordinate attention mechanism is used to better capture location information and channel feature information, highlighting the feature information of small targets and occluded human keypoints. Besides, the parallel cross-correlation convolution is further modified to enhance the capability of feature information extraction. The method's performance is verified on the general pose estimation dataset (COCO) and the dedicated infant pose estimation dataset (SyRIP). The results show that, with a calculation volume (GFLOPs) of only 0.96, the method achieves average accuracies of 73.5% and 91.0% on COCO and SyRIP datasets, respectively, proving that it can significantly reduce the quantity of model parameters and calculation volume without sacrificing pose estimation accuracy. The proposed lightweight estimation model is expected to be deployed on home appliances such as smart terminals, thereby realizing intelligent estimation of abnormal infant poses in home scenarios.

Keywords: infant pose estimation; lightweight; motion; MobileNetV3

前 言

异常的运动姿态是婴幼儿神经肌肉系统疾病(如脑瘫)的主要临床症状[1],后续康复疗效在很大程度上取决于对婴儿早期异常运动姿态的有效评估[2]。现有研究通过使用带有标记的跟踪系统[3]、表面肌电图[4]和角度传感器[5]等方式实现对婴幼儿运动姿态的客观测量,但是这些方法需要在婴儿的身体上安装标记或传感器,这种接触式的测量方式可能会阻

【收稿日期】2024-07-24
【基金项目】国家自然科学基金(32460238);江西省自然科学基金(20232BAB206134)
【作者简介】万金亮,硕士研究生,研究方向:人体姿态估计、智能康复,E-mail: 483839061@qq.com
【通信作者】熊启亮,博士,副教授,研究方向:生物医学信号检测与处理、运动康复、人体姿态估计,E-mail: 70898@nchu.edu.cn

碍他们的自然运动状态，严重情况下甚至会导致婴幼儿拒绝配合。为此，如何对婴幼儿早期异常姿态进行“非接触式”的测量成为一个亟待解决的临床康复工程问题。基于深度学习网络的人体姿态估计是一种从输入视频或连续图像中找到人体信息，并恢复身体主干和关节特定位置的方法，这种方法为获取和分析婴儿非接触式运动信息提供可能[6]。

目前，基于深度学习网络的姿态估计方法主要以回归方法[7]和热图方法[8]为主，其中，基于回归的姿态估计方法直接通过深度神经网络的输出层对人体关键点的坐标进行回归。但是这种方式可能导致对关键点的空间信息的丢失，使得训练模型缺乏空间泛化能力。基于热图的姿态估计方法则是将人体姿态估计从坐标回归问题转化为检测问题，最大程度地保留关键点坐标的空间信息，极大提高学习姿态估计模型的空间泛化能力，从而提高姿态估计的准确性。与基于回归的方法相比，基于热图的姿态估计方法往往能够实现更高的准确性，因此在人体姿态估计领域得到广泛使用（如AggPose[9]）。然而，这些模型普遍网络结构复杂性高、计算需求大，因此很难部署到嵌入式平台和家用设备中。为婴儿的异常运动姿态评估提供一种面向家庭环境的测评方法，尤其是支持父母仅使用智能手机等移动设备即可完成评估，对于降低成本和缓解医疗资源压力具有重要的实际意义。因此，本文旨在开发一种轻量级婴幼儿姿态估计方法，在保证模型准确性的同时，减少模型参数和计算负载，从而进一步促进婴幼儿姿态估计方法在移动设备或嵌入式智能终端中的应用。

1 人体姿态估计研究进展

1.1 姿态估计概念

人体姿态估计也称为人体姿态识别，指利用计算机视觉和机器学习技术从图像或视频中自动推断人体姿态信息，包括关节位置、角度和空间姿态。目前，主流的姿态估计网络模型包括hourglass网络（SHN）[10]、级联金字塔网络（CPN）[11]、SimpleBaseline[8]和用于视觉识别的深度高分辨率表示深度学习网络（HRNet）[12]。其中，SHN利用大量残差模块有效提取多尺度特征，并整合来自不同尺度的特征信息，以获取相关信息，从而有助于识别关键点。通过在特征图上重复执行由高到低分辨率的编码和解码操作，SHN可以逐渐重建出更高分辨率和更详细的姿态特征图[13]。CPN网络主要以级联方式优化姿态估计结果，首先，初始姿态估计网络生成初步的关键点预测。然后，这些预测作为另一个更高分辨率特征金字塔的输入，与原始图像特征融合。随后，融合的特征金字塔被送入下一个级联网络用于进一步关键点预测。通过级联网络的串行操作，CPN可以逐渐改进和校正姿态估计结果，以获得更准确的关键点位置。SimpleBaseline模型使用ResNet作为骨干网络，并仅在ResNet的最后一层堆叠3个反卷积层，以输出高分辨率特征图和关键点热图。该模型通过反卷积层有效提高关键点热图的分辨率，从而增强姿态估计任务的性能。HRNet采用保留不同尺度特征信息的多尺度融合策略，避免传统网络中不同分辨率下的信息丢失。因此，HRNet展现出良好的跨尺度特征捕获能力，从而提高网络对多个层次物体特征的感知敏锐度。此外，HRNet有效地利用并行连接的能力，实现来自不同层的特征的并行处理，并在它们之间建立相互连接的路径。因此，这种策略促进不同层次信息的无缝交流和知识交换，从而改善来自不同层次的信息整合。

1.2 轻量级姿态估计方法

目前主流轻量级姿态估计网络主要为ShuffleNet[14]和HRNet的轻量化版本[15]。其中，ShuffleNetV1在分组卷积之后执行通道混洗，以促进信息传递并提高模型性能。ShuffleNetV2通过去除大量的分组卷积和引入通道分离来减少内存使用。该方法涉及将输入特征图沿通道维度分为两个分支，在每个分支上应用不同的卷积操作，然后将这些操作的结果连接起来，随后进行通道混洗过程。Lite-HRNet[16]是HRNet的轻量化改进网络，它将Shuffle块应用到HRNet，并引入一种高效的条件通道加权单元，以替代Shuffle块中昂贵的1×1卷积。值得注意的是，关于轻量级姿态估计的最新研究主要集中在Lite-HRNet上，并取得令人鼓舞的成果[17]。

为了进一步降低模型参数量和计算量，人体姿态估计领域的研究人员对网络架构也进行针对性修改，这些修改包括移除全连接层，并引入解码部分等。其中，SimpleBaseline利用多个转置卷积层生成高分辨率特征图的方式获得出色的估计精度。虽然模型结构简单，但引入转置卷积带来大量参数和计算成本，导致上采样部分占用大部分计算资源，进而使得网络模型很难部署于算力有限的小型设备。为此，Niu等[18]提出一种高效的上采样模块，显著减少参数数量并实现显著的性能改进，从而在小型设备上实现更快的推理速度和更低的计算成本。PixelShuffle的整体思想是将输入特征图分成重叠的小块，并对每个小块执行像素重新排列[19]。与其他上采样模块相比，PixelShuffle不仅扩展图像分辨率，还允许减少输入特征图通道的数量，显著降低网络的计算成本。

1.3 婴幼儿姿态估计

近年来，随着计算机视觉和深度学习技术的发展，研究人员开始尝试面向婴幼儿姿态估计问题展开相关工作。例如，Reich等[20]首次使用OpenPose从婴儿运动视频中估计出婴儿骨架，并应用浅层多层神经网络对婴儿动作进行分类。Migliorelli等[21]采用由检测卷积神经网络（Convolutional Neural Network, CNN）和回归CNN组成的深度学习框架来定位婴儿关节。Sakkos等[22]使用一维CNN架构从婴儿视频中对婴儿动作进行分类。Geng等[23]使用具有多分支结构的HRNet网络进行骨骼关键识别，以提高预测关键点位置的定位质量。Huang等[24]提出一种自适应模型FIDIP用于稳健的婴儿姿态估计。这些研究使用预训练的成人姿态估计模型作为起点，并在婴儿姿态数据集上进行微调，以适应婴儿姿态的特点。Li等[25]引入一种有效的基于回归的人体姿态识别方法，通过建立级联Transformer来解决儿童身体部分遮挡的问题。为了恢复受损区域的合理人体部位，Zhao等[26]提出两种不同类型的学习先验方法来补偿姿态估计图像中受损区域，进一步能够从被破坏的区域中恢复出具有合理人体形状的人体部位。上述婴儿姿态估计网络涉及计算密集型的深度学习模型（即具有数千万可训练参数），需要大量的计算、内存和能源资源，这限制其在手机等计算能力有限的边缘设备和移动终端上的适用性。

MobileNet模型是Google针对手机等嵌入式设备提出的一种轻量化深度神经网络，使用的核心思想是深度可分离卷积，与Lite-HRNet和ShuffleNetV2[27]相比，MobileNet的最新版本MobileNetV3[28]（图1）引入通道注意模块来增强特征提取，并在瓶颈模块（Bottlenek）的后半部分使用HardSwish[29]替换ReLU激活函数。这种结构在具有更少计算和参数的情况下实现更好的性能[30]。因此，本文选择MobileNetV3作为婴幼儿“轻量化”姿态估计模型的主干网络。

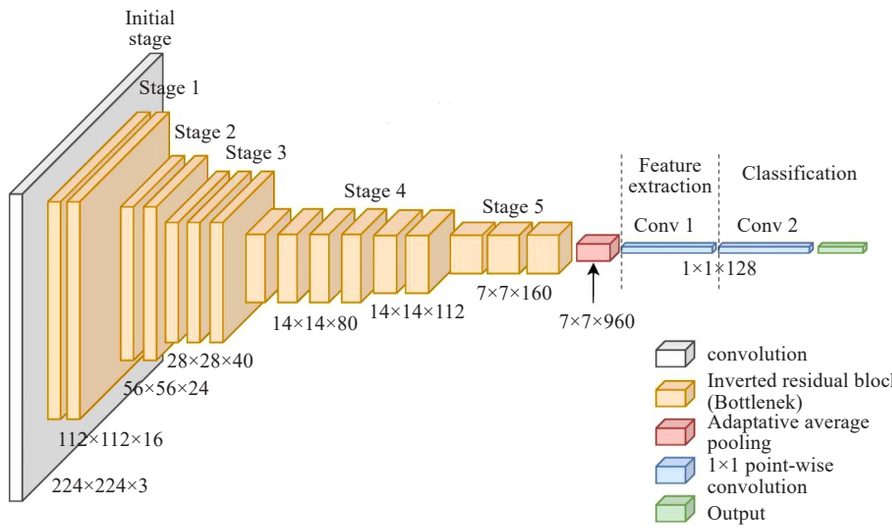


图1 MobileNetV3网络的整体结构示意图

Figure 1 Overall structure of MobileNetV3 network

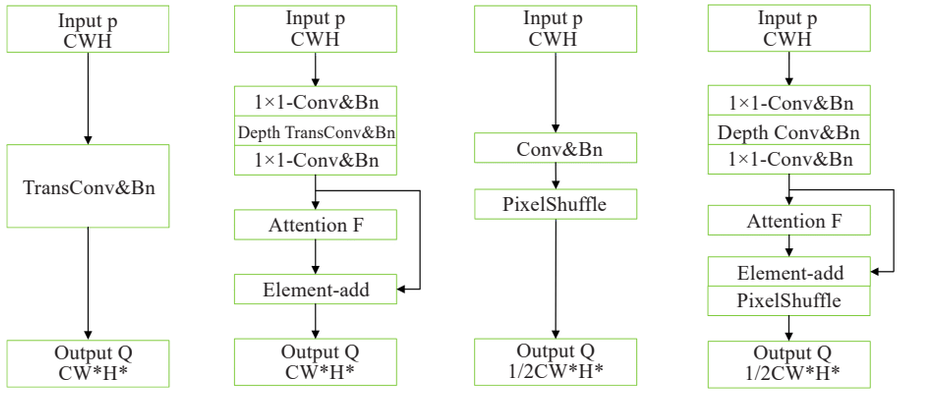
2 方 法

2.1 上采样模块

为了降低模型的参数和计算量，本文基于深度可分离和像素重排原理，设计一种密集上采样深度可分离卷积（DS-PixelShuffle, DSP）以替换MobileNetV3主干网络中的转置卷积。首先，输入卷积层经过MobileNetV3瓶颈模块的深度可分离操作。随后，采用PixelShuffle技术来降低输入卷积层的通道数，同时增加特征图的大小。与转置卷积（图2a）和深度可分离转置卷积（图2b）相比，密集上采样深度可分离卷积（图2d）的创新之处在于其对输入的初始通道扩展，从而消除转置矩阵引入的大量参数和计算负担。与密集上采样卷积（图2c）相比，卷积层被深度可分离卷积取代，进一步减少计算资源的利用。

2.2 注意力模块

在计算机视觉中能够把注意力聚集在图像重要区域的方法被称作是注意力机制。MobileNetV3内置的通道注意力机制只关注卷积层中各个通道的权重而忽视在空间上特征图的相互关系，目前主流的



a:转置卷积　　b:深度可分离转置卷积　　c:密集上采样卷积　　d:密集上采样深度可分离卷积

图2 不同上采样模型结构示意图

Figure 2 Structures of different up-sampling models

兼顾通道和空间的注意力机制往往需要较大的参数和计算量,不适宜部署在终端设备上。基于网络模型轻量化的需求,本文进一步使用坐标注意力(Coordination Attention, CA)模块,如图3所示。该模块以更少的参数,既关注通道信息,又兼顾宽度和高度上的空间信息,可以增强MobileNetV3上Block模块上敏感区域的特征提取能力,有利于提高后续对人体关键点的估计精度。

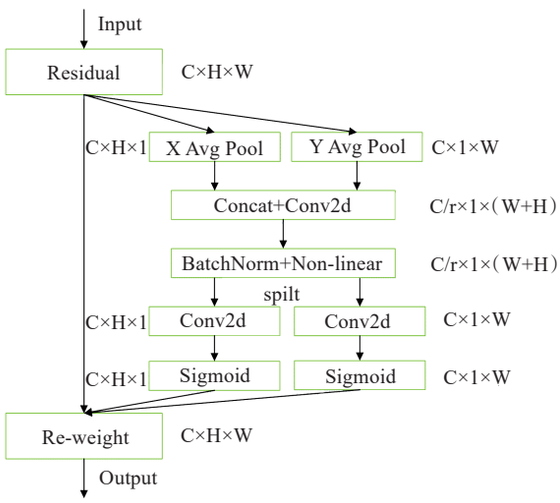


图3 CA网络结构图

Figure 3 Overall structure of CA network

2.3 基于多尺度特征融合的模型结构

人体姿态估计涉及检测关键点的位置和分类不同的关键点,因此人体姿态估计模型要求其网络既具有分类任务的位置不变性,又具有检测任务的位置敏感性。虽然DSP模块本身包含卷积块,但上采样模块的轻量级特性并不能提供足够的参数来完全恢复输入图像的特征信息。因此在解码过程中,上采样可能会丢失感兴趣区域的位置敏感性。为解决这一问题,考虑到浅层高分辨率卷积包含丰富的空间信息,并表现出更强的位置敏感性,本文在MobileNetV3骨干网络的第3、6和12个瓶颈模块,分别通过不同的特征提取方法(a、b、c、d、e)生成3种不同尺度的特征信息。第3、6和12个瓶颈模块通过不同的特征提取方式生成96×72、48×36、24×18这3种不同分辨率的特征图,再对第3、6和12个瓶颈模块生成的相同分辨率的特征图进行融合之后,应用DSP上采样将它们恢复到原始图像分辨率的四分之一。最后,通过1×1卷积处理连接的特征层,分别输出不同位置的关键点位置。本文与HRNet并行连接不同的地方在于使用MobileNetV3集成的Block模块替换常规卷积降低参数,使用DSP上采样替换双线性插值还原特征信息,以保证在网络有效性和计算复杂度上进行平衡。模型的网络结构如图4所示。

3 实 验

3.1 数据集

由微软亚洲研究院发布的COCO数据集是目前应用最广泛的人体姿态估计公共数据集之一[31]。该数据集包含约20万张图片,约有25万个人体实例。每个人物标注了全身的17个关键点,并区分关键点的左右以及是否可见。在本文中,使用该数据集的"train2017"子集作为训练集,该子集包含约5.7万张图片,约有15万个成人人体实例。"val2017"子集作为验证集,包含约5 000张图片,含有约1.8万个成人人体实例。

鉴于本文的主要工作是提出一个面向婴幼儿姿

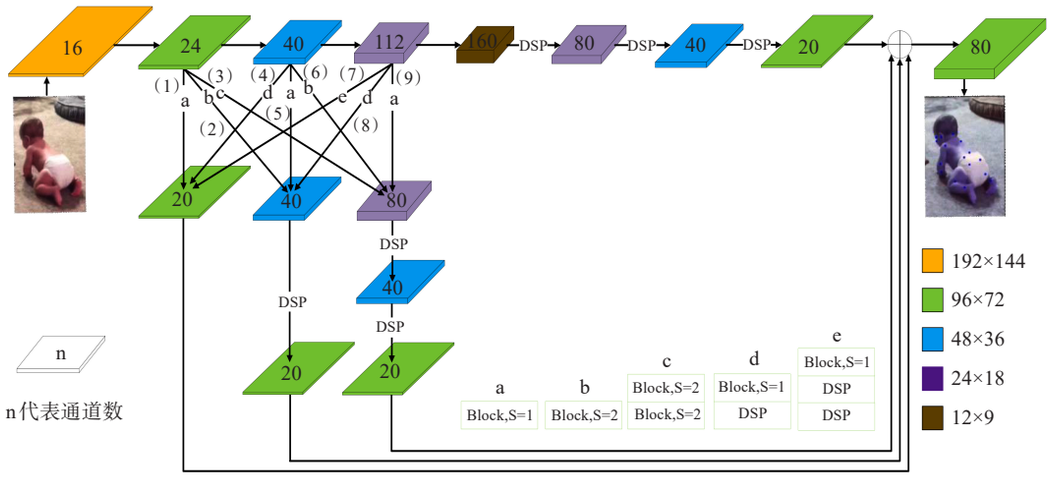


图4 基于MobileNetV3和并行多尺度特征融合的婴幼儿姿态估计网络结构示意图

Figure 4 Infant pose estimation network based on MobileNetV3 and parallel multi-scale feature fusion module

态估计的轻量级网络，因此本文除了COCO数据集验证之外，还在特定婴幼儿姿态估计数据集SyRIP中测试本研究提出网络的性能[28]。SyRIP数据集来源于YouTube和Google Images等平台，包含多样化的真实和合成婴儿图像，该数据集也对婴幼儿身体的17个关键点进行注释。本文将SyRIP数据集的"train_infant"子集（包含约2 000张图像）作为训练集，而"validate_infant"子集（约有100张图像）作为验证集。

3.2 实验设置

本文提出的网络模型使用NVIDIA 1080TI GPU进行训练。操作系统采用的是Ubuntu 18.04，深度学习框架采用的是PyTorch。输入图像的分辨率为384×288，训练过程中应用常见的数据增强技术，如随机缩放、水平翻转和随机裁剪。源代码框架采用基于DarkPose热图解码方法的"human-pose-pytorch"[32]。在模型的训练过程中，设置总共210个epochs，即所有训练样本被训练210次。在实验中，使用Adam优化算法迭代网络模型参数。初始学习率设为0.001，动量参数设为0.9。

3.3 评估指标

本文使用的是COCO数据集中人体姿态估计任务的评估指标-物体关键点相似度（OKS）：

$$\mathrm{OKS}=\frac{\sum_i \exp\left(-\frac{d_i^2}{2s^2k_i^2}\right)\delta(v_i>0)}{\sum_i \delta(v_i>0)} \tag{1}$$

其中，d代表预测关键点坐标和真实关键点坐标之间的欧氏距离；s代表物体尺度因子；k代表归一化因子；v_i代表真实关键点的可见性标志。在COCO数据集中进行计算可以得到以下指标：平均精度（AP）、AP.5和AP.75（当OKS分别为0.5和0.75时的精度）、平均召回率（AR）、AR.5和AR.75（当OKS分别为0.5和0.75时的召回率）。卷积运算的计算复杂度（GFLOPs）计算公式如下所示：

$$\mathrm{GFLOPs}=(2\times D_{\mathrm{in}}\times S^2-1)\times H_{\mathrm{out}}\times W_{\mathrm{out}}\times D_{\mathrm{out}} \tag{2}$$

卷积操作的参数数量（Params）计算公式如下：

$$\mathrm{Params}=(2\times D_{\mathrm{in}}\times S^2-1)\times D_{\mathrm{out}} \tag{3}$$

其中，D_{in}是输入特征通道的数量；S是卷积核的大小；H_{out}、W_{out}、D_{out}分别表示输出特征的高度、宽度和通道数。

4 结果与分析

4.1 模型性能的分析

表1和表2分别展示本文提出的婴幼儿轻量级姿态估计方法与其他姿态估计模型（包括Scnet[33]、MSPN[34]和MMPOSE库中的5种轻量级姿态模型）在COCO（表1）和SyRIP（表2）数据集上的性能对比。例如，在COCO数据集的实验结果表明，与Simple Baseline、Scnet、Hourglass、HRNet和4×MSPN等5个主流姿态估计模型相比，本文提出的轻量级姿态估计方法在模型参数量上分别减少92.0%、92.4%、80.2%、95.9%、97.8%，在计算复杂度上分别减少95.2%、91.9%、93.2%、97.0%、98.2%，而AP分别只降低1.5%、1.7%、0.8%、3.2%、2.9%。在对比主流人体姿态估计模型性能的基础上，本文进一步将提出的婴幼儿轻量级姿态估计方法与Shuffle和MobileNet[35]等几种主流轻量级姿态估计模型进行性能比较。实验结果表明，与Shuffv1、Shuffv2、MobileNetV2等轻量级模型相比，本文提出的方法在参数量上分别减少62.9%、65.9%、73.1%，在计算复杂

度上分别减少68.5%、68.8%、73.1%，并且AP分别提高10.9%、9.7%、5.8%。此外，即便是与目前公认性能优越的轻量级模型Lite-HRNet18和Lite-HRNet30相比，本文方法的AP分别提高5.9%和3.6%。

表1 在COCO验证集上的比较结果

Table 1 Comparative results on COCO validation set

方法	Params/M	GFLOPs	AP/%	AP.5/%	AP.75/%	AR/%	AR.5/%	AR.75/%
Simple Baseline[8]	32.42	20.23	75.0	90.8	82.1	80.0	94.2	86.5
Scnet101[33]	34.01	11.94	75.2	90.6	82.3	80.4	94.3	86.1
Hourglass(256×256)[10]	13.02	14.30	74.3	92.5	81.5	77.3	93.0	83.7
HRNet[12]	63.59	32.88	76.7	91.1	83.2	81.7	94.7	87.5
4×MSPN[34]	120.18	53.60	76.4	90.6	83.4	72.8	82.9	82.6
Shuffv1[14]	6.94	3.05	62.6	86.1	69.5	68.6	90.3	75.3
Shuffv2[27]	7.55	3.08	63.8	86.6	70.7	69.9	91.0	76.5
MobileNetV2[35]	9.57	3.57	67.7	88.2	74.5	73.4	92.0	79.7
Lite-HRNet18[16]	1.13	0.60	67.6	87.7	74.6	73.4	92.0	79.7
Lite-HRNet30[16]	1.76	0.95	69.9	88.4	77.5	75.8	92.7	82.7
本文方法	2.57	0.96	73.5	91.6	80.4	76.3	93.0	82.1

表2 在SyRIP验证集上的比较结果

Table 2 Comparative results on SyRIP validation set

方法	Params/M	GFLOPs	AP/%	AP.5/%	AP.75/%	AR/%	AR.5/%	AR.75/%
Simple Baseline[8]	32.42	20.23	90.1	98.5	98.5	91.6	99.0	98.0
Scnet101[33]	34.01	11.94	51.9	88.0	55.4	55.9	90.0	61.0
Hourglass(256×256)[10]	13.02	14.30	91.0	98.0	98.6	91.7	98.0	99.0
HRNet[12]	63.59	32.88	92.4	98.4	98.4	93.8	99.0	99.0
4×MSPN[34]	120.18	53.60	91.6	98.5	98.5	92.8	98.0	99.0
Shuffv1[14]	6.94	3.05	83.0	99.0	92.0	84.7	99.0	94.0
Shuffv2[27]	7.55	3.08	84.2	100.0	92.3	88.8	100.0	94.0
mobileNetV2[35]	9.57	3.57	86.8	100.0	95.7	88.5	100.0	96.0
Lite-HRNet18[16]	1.13	0.60	88.1	100.0	96.6	89.3	100.0	97.0
Lite-HRNet30[16]	1.76	0.95	89.5	100.0	97.3	91.2	100.0	98.0
本文方法	2.57	0.96	91.0	98.3	98.3	92.9	99.0	99.0

4.2 DSP模块的有效性分析

为了评估DSP模块在参数减少和计算复杂度方面的有效性，本文在MobileNetV3网络上移除了1×1卷积和全连接层，通过串联不同的上采样层探究DSP模块的有效性，表3对比了DSP模块与转置卷积、深度可分离转置卷积和密集上采样卷积方法在COCO和SyRIP数据集上的性能差异。例如，在COCO数据集上的实验结果表明，与传统的转置卷积方法相比，DSP模块成功地将参数数量减少18.7%，并且计算复杂度减少86.6%。此外，这种模型架构的优化还有助于模型精度的提升，在COCO验证集上AP提高1.4%，在SyRIP验证集上AP提高2.8%。将DSP模块与深度可分离转置卷积进行比较时，可以观察到前者在使用的参数数量上略微增加5%，同时计算复杂度大幅减少49.5%。与密集上采样卷积方法相比，DSP模块在使用的参数数量上减少2.3%，计算复杂度减少8.7%。特别是在COCO验证集上，DSP模块的AP达到了68.6%，在SyRIP验证集上更进一步提

高到88.5%。上述结果表明，本文提出的DSP模块不仅实现了与现有上采样模块相当的识别准确性性能，而且在参数和计算复杂度方面都显著减少。

表3 不同上采样模块在COCO和SyRIP数据集上的性能比较

Table 3 Performance comparison among different up-sampling modules on COCO and SyRIP datasets

数据集	方法	Params/M	GFLOPs	AP/%	AP.5/%	AP.75/%	AR/%	AR.5/%	AR.75/%
COCO	转置卷积	3.74	3.90	67.2	89.4	73.9	70.5	90.3	76.4
	深度可分离转置卷积	2.88	1.03	68.6	90.4	76.0	71.8	91.3	78.3
	密集上采样卷积	3.11	0.57	68.6	90.4	75.3	71.8	91.2	78.0
	DSP	3.04	0.52	68.6	90.4	76.1	71.8	91.4	78.3
SyRIP	转置卷积	3.74	3.90	85.7	92.6	92.5	88.0	99.0	95.0
	深度可分离转置卷积	2.88	1.03	86.5	97.3	92.2	89.0	98.0	95.0
	密集上采样卷积	3.11	0.57	88.5	98.8	95.5	89.9	99.0	96.0
	DSP	3.04	0.52	88.5	98.5	95.9	90.6	99.0	97.0

4.3 CA模块的有效性分析

为了评估本文提出的CA注意力模块的有效性，在MobileNetV3网络上移除了1×1卷积和全连接层，串联DSP模块的基础上把传统注意力（SE）修改为CA注意力，并与SE注意力模块进行对比实验。表4展示了CA注意力模块与SE模块在COCO和SyRIP数据集上的性能差异。实验表明，与使用SE注意力模块相比，采用CA注意力模块在降低模型参数量的同时，在COCO数据集和SyRIP数据集上的AP分别提升0.9%和0.5%。

表4 注意力模块的对比实验

Table 4 Comparative experiments of attention modules

数据集	注意力模块	Params/M	GFLOPs	AP/%	AP.5/%	AP.75/%	AR/%	AR.5/%	AR.75/%
COCO	SE	3.04	0.52	68.6	90.4	76.1	71.8	91.4	78.3
	CA	2.01	0.53	69.5	90.6	78.4	72.3	92.0	79.1
SyRIP	SE	3.04	0.52	88.5	98.5	95.9	90.6	99.0	96.0
	CA	2.01	0.53	89.0	98.7	96.3	90.9	99.0	97.0

4.4 并行多尺度特征融合结构的有效性分析

为了验证本文方法中并行多尺度特征融合结构的有效性，并行连接结构如图4所示，本文对如图4所示的并行连接结构中的每个分支进行对比实验。其中，在MobileNetV3网络上移除了1×1卷积和全连接层，并且串联DSP模块的基础上，移除MobileNetV3 SE注意力模块。首先分析并联第3个瓶颈模块生成的3个不同特征图对网络结构的影响，然后分析并联第6、12个瓶颈模块生成的3个不同特征图对网络结构的影响。表5展示了上述不同分支结构在COCO数据集上的对比结果。由表5可知，与并行1、2、3分支相比，并行4、5、6分支的参数量上升1.5%，计算量没有发生显著变化，AP上升0.2%，并行7、8、9分支的参数量上升18.0%，计算量上升9%，AP上升0.5%。上述结果表明，在3个并行的分支里融合更深层的瓶颈模块较浅层的瓶颈模块虽然会略微提升模型的复杂度，但是对整体姿态识别的效果却有明显的提升。因为更深层的卷积往往拥有更为丰富的语义信息，对同一分辨率的特征图，并联更深层的卷积带来的尺度信息有助于更好地理解姿态估计任务中不同关键点的位置关联信息，从而提升关键点估计精度。

4.5 消融实验

为了进一步验证本文提出的CA注意力机制，以及并行多尺度特征融合模块之间的交互效果，基于前文所述的相同硬件和软件环境，在COCO数据集

表5 并行不同分支的网络结构在COCO数据集上的对比实验

Table 5 Comparative experiments of parallel network structures with different branches on COCO dataset

不同分支	Params/M	GFLOPs	AP/%	AP.5/%	AP.75/%	AR/%	AR.5/%	AR.75/%
1、2、3	1.83	0.71	69.5	90.5	76.4	72.4	91.3	78.6
4、5、6	1.86	0.71	69.7	90.5	76.3	72.7	91.5	78.6
7、8、9	2.14	0.77	70.0	90.5	77.2	72.8	91.4	79.1
1~9	2.21	0.95	70.4	90.5	77.2	72.1	91.5	79.3

上进行消融实验。图5所示的单纯主干是利用MobileNetV3骨干网的第1阶段到第5阶段作为编码层，并采用一个3层的DSP模块作为解码层。该配置连接到一个1×1卷积网络，生成17个关键点。另一方面，在主干基础上增加的并行多尺度特征融合分支结构则主要是将MobileNetV3骨干网的第1阶段和第2阶段到第4阶段产生的密集交叉模块整合在一起，生成3个分支。每个分支经历DSP上采样，然后合并到一起，并连接到一个1×1卷积网络。

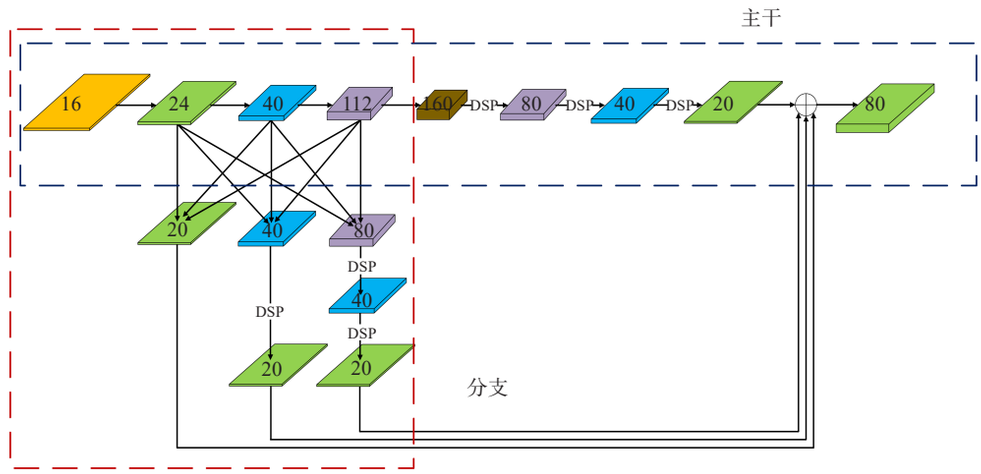


图5 消融实验示意图

Figure 5 Diagram of ablation study

表6展示在COCO数据集上分别进行消融实验的结果，其中重点对比是否加入并行多尺度特征融合的分支结构和CA注意力模块对于模型性能的影响。首先，实验中使用主干网络输出关键点坐标。其中使用MobileNetV3内置的SE注意力模块。然后，在不改变SE注意力模块的基础上，增加并行多尺度特征融合分支。最后，在并行多尺度特征融合结构的基础上把SE注意力模块替换为CA注意力模块。结果显示，与单纯使用主干网络结构相比（表6第1行），引入多尺度特征融合分支结构（表6第2行）以后AP提升3.1%，在此基础上，进一步将SE注意力模块替换为CA注意力模块（表6第3行）以后，AP进一步提升1.8%。上述结果表明引入并行多尺度特征融合结构对整体姿态的识别有显著作用，对不同卷积层的交叉连接促进各个尺度的信息进行交互，能更好地理解各个关键点之间的位置关系。同时，CA注意力模块的使用能够有效地弥补卷积过程中由浅层特征可能引起的信息损失，从而提高关键点检测任务中的位置灵敏度。

4.6 定性分析

本文在婴幼儿专用姿态估计数据集SyRIP测试集中选取家庭环境中不同场景下的不同婴幼儿姿态样例进行可视化，如图6所示。无论是躺、爬行姿态，还是匍匐姿态，本文提出的方法在SyRIP数据集上展示出良好的关键点估计能力。即使面临单侧爬行姿态，基于并行多尺度特征融合的方法也能有效捕获图片全局信息，这也说明本文提出的面向婴幼儿的轻量级姿态估计方法在不同姿态和多角度变化上的可靠性和稳定性。

表6 COCO数据集上的消融实验结果

Table 6 Results of ablation study on COCO database

是否采用多尺度特征融合分支结构	是否采用CA注意力模块	Params/M	GFLOPs	AP/%	AP.5/%	AP.75/%	AR/%	AR.5/%	AR.75/%
×	×	3.04	0.52	68.6	90.4	76.1	71.8	91.4	78.3
√	×	3.92	0.95	71.7	90.6	78.5	74.5	91.8	80.7
√	√	2.57	0.96	73.5	91.6	80.4	76.3	93.0	82.1

图6 本文方法在SyRIP数据集上所检测到的骨骼关键点效果

Figure 6 Bone key points detected by the proposed method on SyRIP dataset

5 结 论

为了满足在计算能力有限的家用移动终端上部署婴幼儿姿态估计模型的应用需求，本文以轻量级网络MobileNetV3作为编码的骨干，提出一种婴幼儿轻量级姿态估计方法，该方法在通用数据集COCO和婴幼儿数据集SyRIP上进行的实验结果，均表明其在确保姿态估计准确性的同时，能够显著减少模型参数和计算复杂度。与其他人体姿态估计任务相比，本文是首次尝试探索轻量级的婴幼儿姿态估计方法，希望本文相关工作可以为面向家庭环境中的婴幼儿姿态估计应用提供启发和借鉴。

【参考文献】

[1] Simsek C, Mengi A, Yalcinkaya EY. The effect of psychodrama on quality of life and sleep in mothers of children with cerebral palsy[J]. Arts Psychother, 2021, 72: 101726.

[2] Chiarello LA, Palisano RJ, Avery L, et al. Longitudinal trajectories and reference percentiles for participation in family and recreational activities of children with cerebral palsy[J]. Phys Occup Ther Pediatr, 2021, 41(1): 18-37.

[3] Kolahi A, Hoviattalab M, Rezaeian T, et al. Design of a marker-based human motion tracking system[J]. Biomed Signal Process Control, 2007, 2(1): 59-67.

[4] Day S. Important factors in surface EMG measurement[EB/OL]. http://www. bortec. ca/Images/pdf/EMG% 20 measurement% 20 and% 20recording.pdf.

[5] Lim CK, Luo ZQ, Chen IM, et al. A low cost wearable optical-based goniometer for human joint monitoring[J]. Front Mech Eng, 2011, 6 (1): 13-22.

[6] Li XY, Liu YH, Gao ZJ, et al. Computer vision online measurement of shiitake mushroom (Lentinus edodes) surface wrinkling and shrinkage during hot air drying with humidity control[J]. J Food Eng, 2021, 292: 110253.

[7] ToshevA, SzegedyC. DeepPose: human pose estimation *via* deep neural networks[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE, 2014: 1653-1660.

[8] Xiao B, Wu HP, Wei YC. Simple baselines for human pose estimation and tracking[C]//Computer Vision-ECCV 2018. Cham: Springer International Publishing, 2018: 472-487.

[9] Cao X, Li XY, Ma LY, et al. AggPose: deep aggregation vision transformer for infant pose estimation[C]//Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence AI for Good. California: IJCAI, 2022: 5045-5051.

[10] Newell A, Yang KY, Deng J. Stacked hourglass networks for human pose estimation[C]//Computer Vision-ECCV 2016. Cham: Springer International Publishing, 2016: 483-499.

[11] Chen YL, Wang ZC, Peng YX, et al. Cascaded pyramid network for multi-person pose estimation[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE, 2018: 7103-7112.

[12] Wang JD, Sun K, Cheng TH, et al. Deep high-resolution representation learning for visual recognition[J]. IEEE Trans Pattern Anal Mach Intell, 2021, 43(10): 3349-3364.

[13] Susanto Y, Livingstone AG, Ng BC, et al. The hourglass model revisited [J]. IEEE Intell Syst, 2020, 35(5): 96-102.

[14] Zhang XY, Zhou XY, Lin MX, et al. ShuffleNet: an extremely efficient convolutional neural network for mobile devices[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE, 2018: 6848-6856.

[15] Howard AG, Zhu ML, Chen B, et al. MobileNets: efficient convolutional neural networks for mobile vision applications[EB/OL].

(2017-04-17). https://arxiv.org/abs/1704.04861.
[16] Yu CQ, Xiao B, Gao CX, et al. Lite-HRNet: a lightweight high-resolution network[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2021: 10435-10445.
[17] Li Q, Zhang ZY, Xiao F, et al. Dite-HRNet: dynamic lightweight high-resolution network for human pose estimation[C]//Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence Main Track. California: IJCAI, 2022: 1095-1101.
[18] Niu SL, Ou WH, Feng SH, et al. Designing compact convolutional filters for lightweight human pose estimation[J]. Wirel Commun Mob Comput, 2021, 2021(1): 1333250.
[19] Shi WZ, Caballero J, Huszár F, et al. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2016: 1874-1883.
[20] Reich S, Zhang DJ, Kulvicius T, et al. Novel AI driven approach to classify infant motor functions[J]. Sci Rep, 2021, 11(1): 9888.
[21] Migliorelli L, Moccia S, Pietrini R, et al. The babyPose dataset[J]. Data Brief, 2020, 33: 106329.
[22] Sakkos D, Mccay KD, Marcroft C, et al. Identification of abnormal movements in infants: a deep neural network for body part-based prediction of cerebral palsy[J]. IEEE Access, 2021, 9: 94281-94292.
[23] Geng ZG, Sun K, Xiao B, et al. Bottom-up human pose estimation *via* disentangled keypoint regression[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2021: 14671-14681.
[24] Huang XF, Fu NH, Liu SJ, et al. Invariant representation learning for infant pose estimation with small data [C]//2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021). Piscataway, NJ, USA: IEEE, 2021: 1-8.
[25] Li KY, Zhang M, Xu MP, et al. Ship detection in SAR images based on feature enhancement Swin transformer and adjacent feature fusion [J]. Remote Sens, 2022, 14(13): 3186.
[26] Zhao ZB, Liu W, Xu YY, et al. Prior based human completion[C]// 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2021: 7947-7957.
[27] Ma NN, Zhang XY, Zheng HT, et al. ShuffleNet V2: practical guidelines for efficient CNN architecture design[C]//Computer Vision-ECCV 2018. Cham: Springer International Publishing, 2018: 122-138.
[28] Abd Elaziz M, Dahou A, Alsaleh NA, et al. Boosting COVID-19 image classification using MobileNetV3 and aquila optimizer algorithm[J]. Entropy (Basel), 2021, 23(11): 1383.
[29] Avenash R, Viswanath P. Semantic segmentation of satellite images using a modified CNN with hard-swish activation function[C]// Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications. Setúbal, Portugal: SciTePress, 2019: 413-420.
[30] Qian SY, Ning CR, Hu YP. MobileNetV3 for image classification[C]// 2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE). Piscataway, NJ, USA: IEEE, 2021: 490-497.
[31] Lin TY, Maire M, Belongie S, et al. Microsoft COCO: common objects in context [C]//Computer Vision-ECCV 2014. Cham: Springer International Publishing, 2014: 740-755.
[32] Zhang F, Zhu XT, Dai HB, et al. Distribution-aware coordinate representation for human pose estimation [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2020: 7091-7100.
[33] Han K, Rezende RS, Ham B, et al. SCNet: learning semantic correspondence [C]//2017 IEEE International Conference on Computer Vision (ICCV). Piscataway, NJ, USA: IEEE, 2017: 1849-1858.
[34] Li WB, Wang ZC, Yin BY, et al. Rethinking on multi-stage networks for human pose estimation[EB/OL]. (2019-05-30). https://arxiv.org/abs/1901.00148.
[35] Sandler M, Howard A, Zhu ML, et al. MobileNetV2: inverted residuals and linear bottlenecks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE, 2018: 4510-4520.

（编辑：陈丽霞）