计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2023,32(7):121−128 [doi: 10.15888/j.cnki.csa.009112]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

轻量型多路特征融合人体姿态估计①

张国有, 高　希

(太原科技大学 计算机科学与技术学院, 太原 030024)
通信作者: 高　希, E-mail: s20202002005@stu.tyust.edu.cn

摘　要: 基于深度学习的人体姿态估计广泛应用于姿态识别、人机交互等领域. 为了提升人体关键点的检测精度, 很多网络采用运算量、参数量和复杂度不断增加的模型架构, 导致无法直接部署到低算力设备. 为了解决上述问题, 本文提出了一种多路特征注意力融合的轻量型方法. 模型基于 HigherHRNet 网络进行轻量化设计和训练, 包括: 采用通道拆分和通道混洗, 解决分组卷积后特征层之间存在的信息隔离; 采用线性运算的特征生成方法, 解决不同特征层之间存在的冗余性; 采用融合注意力信息的方法, 缓解因轻量化导致的准确率下降. 在 MS COCO 数据集上完成了模型的训练、测试、可视化以及消融实验. 实验结果表明本文的轻量化方法在保证直观的检测精度前提下, 能够显著降低人体姿态估计的计算量.

关键词: 轻量型; 特征融合; 注意力特征; 人体姿态估计; 卷积神经网络

引用格式: 张国有,高希.轻量型多路特征融合人体姿态估计.计算机系统应用,2023,32(7):121–128. http://www.c-s-a.org.cn/1003-3254/9112.html

Lightweight Human Pose Estimation Based on Multi-branch Feature Fusion

ZHANG Guo-You, GAO Xi

(College of Computer Science and Technology, Taiyuan University of Science and Technology, Taiyuan 030024, China)

Abstract: Human pose estimation based on deep learning is widely used in pose recognition, human-computer interaction, and other fields. In order to improve the detection accuracy of key points of the human body, many networks adopt a model architecture with increasing calculation amount, parameter amount, and complexity, which is impossible to be directly deployed to low-computing devices. To solve the above issues, this study proposes a lightweight method for multi-branch feature attention fusion. The model is based on the HigherHRNet network for lightweight design and training. Specifically, channel splitting and channel shuffling are adopted to solve the information isolation between feature layers after group convolution; the feature generation method of linear operation is used to address the redundancy between different feature layers; the method of fusing attention information is employed to alleviate the accuracy drop caused by lightweight. The training, testing, visualization, and ablation experiments of the model are completed on the MS COCO dataset. The experimental results show that the lightweight method in this study can significantly reduce the calculation amount of human pose estimation under the premise of ensuring intuitive detection accuracy.

Key words: lightweight; feature fusion; attention feature; human pose estimation; convolutional neural network (CNN)

人体姿态估计的任务是根据图像获取包含于其中的人体关键点信息, 根据关键点位置信息和各个关键点之间的连接信息绘制人体骨骼框架, 在计算机视觉任务中具有充分的挑战性和应用性.

2014 年 Toshev 等人[1] 使用深度学习方法对姿态估计问题提出了 DeepPose 网络. 越来越多的深度学习

① 收稿时间: 2022-11-15; 修改时间: 2022-12-23; 采用时间: 2023-01-06; csa 在线出版时间: 2023-04-28
CNKI 网络首发时间: 2023-05-04

方法应用于姿态估计任务中, 且在实际表现中远超传统方法. 然而, 随着各种姿态估计算法的运算量、参数量、复杂度的不断增加, 姿态估计任务在性能、检测精度方面也不断提高. 如何保持现有较高检测精度的前提下, 降低参数量以及运算量, 以提高网络运行效率并达到实时的姿态检测速率, 是姿态估计任务在追求高精度历程中要考虑并解决的问题.

依据检测人物数量的不同, 可将人体姿态估计分为单人姿态估计和多人姿态估计, 其中多人姿态估计任务中已经涵盖了单人姿态估计. 本文以多人姿态估计作为研究内容.

在多人姿态估计任务中, 存在检测目标人物的数量和分属不同人体的关键点都是不确定的问题, 需要分别完成关键点的检测和分组两个子任务. 根据任务执行的先后顺序不同, 多人姿态估计可分为自顶向下 (top-down) 和自底向上 (bottom-up).

自顶向下算法, 使用目标检测算法检测人体[2−5], 再使用关键点检测网络进行单人姿态估计. 此类算法受限于人体检测算法, 且对图像中所有目标人物循环完成关键点检测, 虽然对关键点的定位精度较高, 但存在检测效率较低、内存消耗较多的问题.

自底向上算法相较于自顶向下算法的区别在于, 不再使用人体检测器检测人体边界框, 而是首先检测所有图像中的关键点, 然后将所有检测到的关键点分组到不同的人体. 此类算法具有检测速率相对较快、内存消耗少的优点, 但针对较难检测的关键点容易出现漏检、误检等问题.

2019 年提出的 HRNet 网络[6] 是一种自顶向下算法, 针对堆叠沙漏网络 (stacked hourglass network, Hourglass)[7] 存在低分辨率特征的信息损失问题, 提出了并联化的高、低分辨率特征融合的网络结构, 重新验证了保持高分辨率特征在关键点检测中的有效性.

2020 年 Cheng 等人[8] 在高分辨率网络 HRNet 基础上, 针对小尺寸与不同尺寸的人体关键点检测进行改进, 提出了自底向上的姿态估计网络 HigherHRNet. 以 HRNet 为主干增加了反卷积模块, 采用多分辨率训练和热图聚合策略, 可预测具有尺度感知的热图, 提高了关节点预测效率且精确度远高于 HRNet, 再根据自底向上算法采用关联嵌入标签算法对关节点进行分组.

Ma 等人[9] 提出的 ShuffleNet V2 网络的基本组成单元, 通过通道分离 (channel split) 方式将特征一分为二, 分别经由各自的深度可分离卷积和逐点卷积完成特征提取, 最终使用通道混洗 (channel shuffle) 方式完成特征融合. Han 等人[10] 对深度神经网络特征图进行分析, 发现相同特征图的不同特征层之间具有较高的冗余性. 从特征层之间冗余和轻量化的角度, 将特征图一部分由常规卷积生成, 另一部分则由新特征图采用线性变化得到. 两种轻量化的思想和网络单元设计, 均在保证网络算法性能前提下大幅降低了网络的参数量和运算量, 为高分辨率网络的轻量化实时运行提供了可能.

近年来, 注意力模型 (attention model) 被广泛应用于深度学习任务中, 具有即插即用、关注相关信息忽略不相关信息、提升模型性能的优点, 是深度学习技术中被广泛关注与研究的核心技术. Zhao 等人[11] 提出的特征金字塔注意力网络 (pyramid feature attention network), 分别采用平均池化和最大池化的双路分支方法, 提取空间注意力和通道注意力, 最终将注意力信息融合到特征图进行自适应更新.

基于上述研究, 本文以 HigherHRNet 存在的高分辨率特征网络的参数量大、计算复杂的问题, 提出了轻量化的多路特征融合的网络 (lightweight multi-branch feature fusion network, L-MFNet), 通过引入各种轻量化方法来达到对高分辨率网络的轻量化, 同时引入注意力机制, 提升网络对关键点信息的提取. 本文的主要工作如下.

(1) 提出了适应高分辨率网络轻量化的 Shuffle 模块和 Ghost 模块, Shuffle 模块应用于网络初始阶段负责提取图像的底层特征信息, 而 Ghost 模块则应用于提取高层语义信息.

(2) 提出一种多路注意力融合模块, 其中包含空间注意力和通道注意力. 通过增加特征图的通道注意力特征层和空间注意力特征层, 增加少量计算从而有效提升网络对关键点信息的提取.

1 相关工作

1.1 轻量化网络

网络轻量化的目的在于, 以降低算法计算量的方式, 使算法在具有较高性能的同时, 提高运行速度. 在人工设计的轻量化网络方面, 包括了基于各种轻量化卷积操作的 MobileNet 系列网络[12−14]、ShuffleNet 系

列网络[9,15], 基于特征图冗余性的 GhostNet 网络[10]. 其中, 轻量化网络的常用操作包括: 将特征层与卷积核的深度拆分的分组卷积 (group convolution)、深度可分离卷积 (depthwise separable convolution)、特征与卷积核的空间拆分的空间可分离卷积 (spatial separable convolution)、最小化卷积核大小的逐点卷积 (pointwise convolution).

本文基于上述轻量化网络算法意在通过引入轻量化模块, 保证高分辨率特征的网络能够实时运行在低算力设备上.

1.2 注意力网络

注意力机制具有关注相关信息而忽略不相关信息的优点, 克服了传统神经网络的一些局限, 例如: 缺乏多特征的提取和强化、严格的网络特征输入顺序. SE-Net 模块[16] 是一种提取通道注意力信息的网络结构. 其依据注意力机制的思想, 通过全局平均池化操作对特征图的空间信息进行压缩, 并连接全连接层对不同通道间的特征层赋值不同权重, 最终使用特征融合方式将注意力信息与特征图融合. DA-Net 网络[17] 采用并联的方式提取通道注意力和空间注意力. 在注意力提取方面, 对特征图进行矩阵操作构建特征图的两种相关性矩阵, 一种矩阵用来表征特征图的通道相关性, 另一种矩阵用于表征特征图的空间相关性. 而金字塔特征注意力网络[11] 则串联提取通道注意力和空间注意力, 并分别采用平均池化和最大池化, 以丰富注意力的表达. 实验表明, 注意力信息的算法在各种任务中有着不错的表现.

2 L-MFNet 网络结构

本文姿态估计网络 L-MFNet 以 HigherHRNet 网络的 w32 版本为基础框架进行轻量化和融合多路注意力改进, 改进方法同时适用于 HigherHRNet 网络的其他版本.

L-MFNet 网络结构如图 1 所示, 其构建关键点提取网络共包含 4 个阶段, 分别为 Stage1、Stage2、Stage3 和 Stage4. Stage1 由 Stem 模块、4 次 Shuffle-Neck 模块串联和 Tansformer 模块构成. 其中, Stem 模块通过串联两层卷积层将图像特征图降低为原图 1/4 分辨率, 获得维度为 64×128×128 的特征图. Transformer 模块通过并联多个卷积层将输入特征图分辨率折半降低, 且在网络的其他阶段皆有应用. Stage2、Stage3、Stage4 根据输入不同分辨率特征图的数量并联相同数量的子网. 这些子网的前向传播顺序为: Attention 模块、Shuffle-Neck 模块、Ghost-Block 模块.

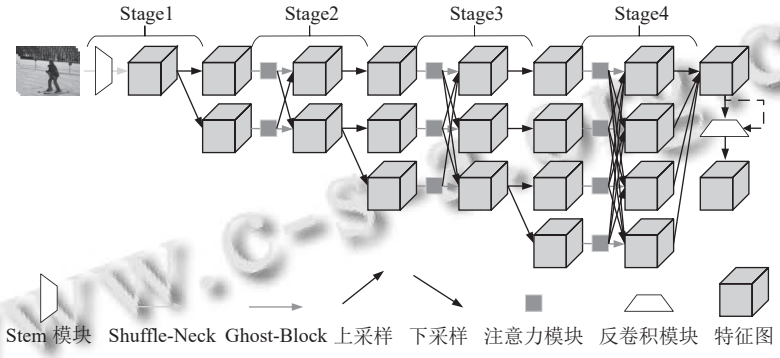

图 1 L-MFNet 网络结构

在网络最后的 Stage4 阶段, 将不同分辨率的特征图通过上采样方法逐级提升, 各支路特征图分辨率, 知道各支路特征图统一到相同的 128×128 分辨率, 然后采用特征融合和串联卷积的方式得到 17×128×128 的关键点热力图 (heatmap). 为了进一步提升关键点热力图的分辨率, 串联具有残差连接的反卷积模块, 进一步将热力图分辨率提升至 256×256, 同时特征图的 17 层分别对应了 17 个人体关键点热力图.

2.1 Shuffle-Neck 模块

Shuffle-Neck 模块的结构如图 2 所示. 首先通过串联两层逐点卷积, 将特征图深度先后调整为 128、64. 将深度为 128 的特征图通过通道拆分 (channel split) 方法, 将特征图在 channel 维度将特征图一分为二, 一半不做任何处理, 另一半串联两层逐点卷积和深度可分离卷积, 用以完成特征图在通道和空间维度的信息交互. 另外, 将深度为 64 的特征图输入 Attention 模块, 得

到空间注意力特征图和通道注意力特征图. 将得到的 4 组特征图完成通道拼接和通道混洗, 在 channel 维度完成特征信息的融合.

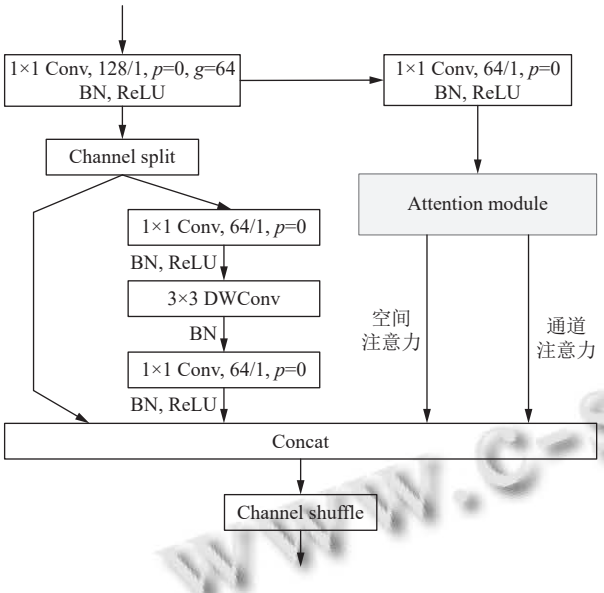

图 2 Shuffle-Neck 模块

分组卷积是轻量化网络的常用手段, 分组数越大参数量越小. 在分组卷积运算生成特征中, 不同特征层之间不参与彼此的卷积运算, 故称为不同特征层的信息隔离. 内存访问量 MAC (输入特征图、输出特征图、卷积核的内存所需空间) 的计算如式 (1) 所示:

$$\begin{aligned} MAC &= hwc_1 + hwc_2 + \frac{c_1c_2}{g} \\ &= hwc_1 + \frac{Bg}{c_1} + \frac{B}{hw} \\ &\geqslant 2\sqrt{hwB} + \frac{Bg}{hw} \end{aligned} \tag{1}$$

其中, h、w 表示输入和输出特征图的宽、高, c_1 表示输入特征图深度, c_2 表示输入特征图深度, B 是值为 hwc_1c_2 的常量.

当卷积层的输入特征与输出特征的深度相等时, MAC 最小. 所以, Shuffle-Neck 模块中使用的卷积层输入特征与输出特征深度相等.

2.2 Ghost-Block 模块

对于相同特征图的不同特征层之间存在的相互冗余和相似关系, 没有必要使用大量冗余的卷积核和运算生成. Ghost module 对于输入特征图 , 利用特征图的各个图层之间的相关性和冗余性, 将传统卷积层划分为两部分. 一部分采用 m 个卷积核的普通卷积运算得到 , 如式 (2) 所示. 另一部分则根据第 1 部分得到的特征图 进行线性运算得到 , 如式 (3) 所示. 最终, 以通道拼接方式将两部分特征图拼接成深度为 n 的特征图.

$$Y' = X \times f' \tag{2}$$

其中, $f' \in R^{c\times k\times k\times m}$的卷积运算, h、w 为输入特征图大小, h'、w'为输入特征图大小.

$$y_{ij} = \Phi_{i,j}(y'_i),\ \forall i = 1,\cdots,m,\ j = 1,\cdots,s \tag{3}$$

其中, s=n–m, y'_1表示特征图Y'_1的第 i 层特征, $\Phi_{i,j}$表示生成第 j 个 Ghost 特征图$y_{i,j}$的第 j 个线性运算.

Ghost-Block 模块的输入特征图与输出特征图具有相同维度, 为了平衡网络的轻量化和算法的有效性, Ghost module 中普通卷积和线性运算在输出特征层数中, 采用相同比例生成. 本文的线性运算采用深度可分离卷积, 当输入特征图深度为 32、大小为 128 时, 根据计算对比可节省 30.1% 计算量. Ghost module 与普通卷积神经网络在参数量方面相比, 仅需较少的普通卷积和线性运算, 同时保持相似的性能, 如图 3 所示.

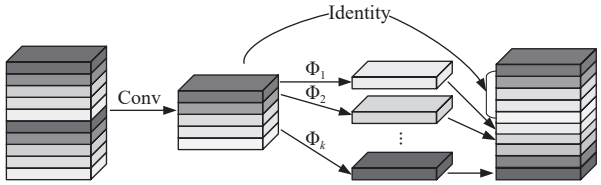

图 3 Ghost module

Ghost-Block 模块包括两个 Ghost module 和 1 个 Attention module, 主路串联两个 Ghost module 后, 将 Attention 模块提取的特征图注意力信息以矩阵相加的方法完成特征融合, 如图 4 所示.

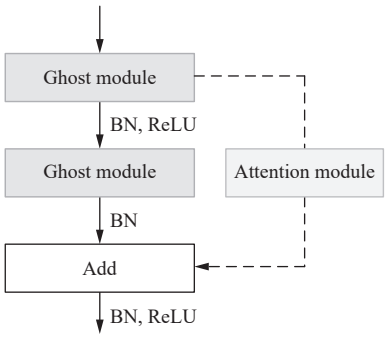

图 4 Ghost-Block 模块

2.3 Attention 模块

Attention 模块利用特征图在空间和通道两种维度蕴含了大量注意力信息的特点, 分别采用平均池化和

最大池化两种方法提取通道注意力信息和空间注意力信息, 其结构如图 5 所示.

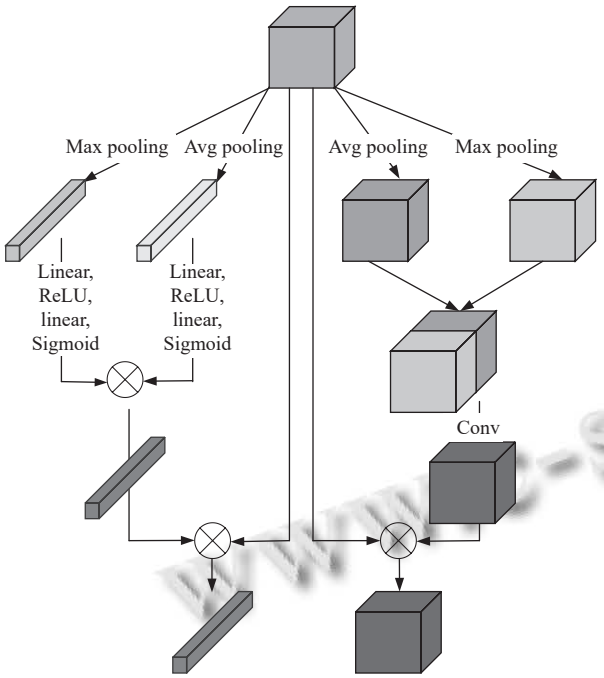

图 5 Attention 模块

通过并行的分支模块对输入特征图提取通道注意力信息和空间注意力信息, 之后使用矩阵相乘方法对输入特征图进行通道和空间的注意力加权, 得到通道和空间注意力特征图.

给定特征图 作为 Attention 模块的输入, 一维特征是特征图 F 的通道注意力信息, 二维特征是特征图 F 的空间注意力信息. 注意力信息的提取与融合过程可概括为:

$$C' = M_C(F) \otimes F \tag{4}$$

$$S' = M_S(F) \otimes F \tag{5}$$

其中, $\otimes$表示维度不等的矩阵乘法. 在计算中, 注意力信息M_C、M_S与 F 采用广播机制完成维度扩展以适应维度不匹配问题. C'和S'是 Attention 模块的最终输出.

3 实验

3.1 实验环境

实验在 Ubuntu 20.04 系统中搭建 PyTorch 1.8 深度学习框架, Python 语言版本为 3.8. 实验硬件使用 32 核心 CPU、32 GB 内存、两块显存为 24 GB 的 RTX3090 显卡, 并搭建了“一机双卡”的分布式训练环境, 以降低实验模型的训练时间. 实验使用 Adam 优化器, 设置初始学习率为 1E–3, 当迭代 MS COCO 数据集 200 周期学习率降低到 1E–4. 为了平衡关键点的检测损失和关键点的分组损失, 赋予其权重分别为 1、1E–3.

3.2 MS COCO 数据集

本实验在 MS COCO (Microsoft common objects in context) 数据集上进行模型的训练与验证. MS COCO 数据集包含超过 250 000 张含有 17 个关键点标签的图像用于人体姿态估计.

3.3 评价指标

为了衡量真实值和预测人体关键点之间的相似度, 采用 MS COCO 数据集官方提供的评价标准 *OKS* (object keypoint similarity) 作为对本文模型评价的度量方法. *OKS* 的定义如式 (6) 所示:

$$OKS = \frac{\sum_i \frac{-d_i^2}{\ell^{2s^2k_i^2}} \delta(v_i > 0)}{\sum_i [\delta(v_i > 0)]} \tag{6}$$

其中, d_i表示标注关键点和预测关键点之间的欧氏距离. s 表示行人的尺度因子, 其值为人体检测面积的平方根, 表达式: $s = \sqrt{wh}$, w、h 人体检测框的宽、高. δ 是关键点的归一化因子, 其值是统计样本中标签值与真实值之间的标准差, 在 MS COCO 数据集中不同类型关键点具有不同的归一化因子, 值越大代表关键点越难以检测. v_i表示关键点的可见类型, 0 表示关键点未标注、1 表示关键点已标注但无遮挡、2 表示关键点已标注但有遮挡.

本文实验使用标准的平均准确率和召回率对实验结果进行分析验证, 包括: mAP (*OKS*=0.5, 0.55, ⋯, 0.9, 0.95 时的平均准确率)、AP50 (*OKS*=0.5 的检测准确率)、AP75 (*OKS*=0.75 的检测准确率)、APM (检测面积在 32–96 的中型尺度人体)、APL (检测面积大于 96 的大型尺度人体)、AR (*OKS* 等于 0.5, 0.55, ⋯, 0.9, 0.95 时的平均召回率). 为了衡量模型在轻量化和准确率方面是否均衡, 将 Params (参数量)、GFLOPs (运算量)、Input size (输入尺寸) 加入到实验结果分析中.

3.4 实验结果

本实验将 L-MFNet 网络在 MS COCO 2017 数据集上分别完成训练和测试, 并与其他性能较好的姿态估计网络在骨干网络、图像大小、参数量、计算量、准确率方面进行对比实验.

表 1 是本文方法在 MS COCO 2017 验证集上获取的实验性能结果和其他模型结果的对比数据, 实验结果表明 L-MFNet 相较于其他网络在相同准确率的情况下, 以更少的参数量和计算量, 获得了相当的准确率. 其中, 对比 HRNet 网络, 本文模型虽然在平均准确率 AP 以及平均召回率 AR 方面分别降低 4.3%、2.7%, 但在模型参数量方面降低 8.6M、在运算复杂度方面 GFLOPs 降低 4.15. 从对比的数据分析不难看出, 本文方法以牺牲少量模型准确率性能为代价, 使得模型的运行效率得到明显提升. 在方法类别中, 由于本方法属于自底向上的姿态估计方法, 相较于自顶向下方法, L-MFNet 网络不仅要完成关键点的检测同时完成关键点的分组任务. 表 1 中的方法均属于自顶向下方法, 且计算量只包含关键点检测而不包含之前的人体检测网络的计算量. 在与同为轻量化网络 Lite-HRNet-30 网络实验结果对比, 在 *OKS* 为 0.5 的平均准确率 AP50 略有降低、参数量提升 1.9M、计算量 GFLOPs 提升 2.64, 其他指标均高于 Lite-HRNet. 本文融合全局注意力信息的方法, 类似于人为观察关键点由全局到局部的过程, 且尺寸越大的人体提取关键点越明显. 同时, 本文方法将输入模型图像分辨率提升 1 倍, 不仅提高模型输出热力图分辨率, 而且提升人体关键点的定位准确度.

表 1 MS COCO 2017 验证集实验结果与其他方法对比

方法	Backbone	Input size	Params ($\times10^7$)	GFLOPs	mAP (%)	AP50 (%)	AP75 (%)	APM (%)	APL (%)	AR (%)
SimpleBaseline[18]	ResNet-50	256×192	3.40	8.90	70.4	88.6	78.3	67.1	77.2	76.3
SimpleBaseline[18]	ResNet-101	256×192	5.30	12.4	71.4	89.3	79.3	68.1	78.1	77.1
SimpleBaseline[18]	ResNet-152	256×192	6.86	15.7	72.0	89.3	79.8	68.7	78.9	77.8
HRNet-32[6]	HRNet-32	256×192	2.85	7.10	73.4	89.5	80.7	70.2	80.1	78.9
Lite-HRNet-18[19]	Lite-HRNet-18	256×192	1.10	0.20	64.8	86.7	73.0	62.1	70.5	71.2
Lite-HRNet-30[19]	Lite-HRNet-30	256×192	1.80	0.31	67.2	88.0	75.0	64.3	73.1	73.3
Ours (MFNet)	MFNet	512×512	1.99	2.95	69.1	86.2	77.0	67.5	76.1	76.1

表 2 为本文方法在 MS COCO 2017 测试集上的实验结果与其他方法在 MS COCO 2017 测试集上的对比数据, 其中 OpenPose、Associative Embedding 和 HigherHRNet 属于自底向上的人体姿态估计算法, 其余则属于自顶向下的人体姿态估计算法. 根据表 2 中的实验数据和对比情况, 本文方法在姿态估计算法的轻量化方法存在很大提升, 且准确率虽然没有与 HigherHRNet 网络相当, 但相对比与 Hourglass、PersonLab 等早期深度学习的姿态估计算法, 在参数量和准确率方面具有较大提升. 较 HigherHRNet-W32, 本文方法在 mAP、AP50、AP75、APM、APL 的准确度度量方面略有下降, 分别降低 1.0%、1.3%、1.3%、0.9%、1.9%. 本文方法较 OpenPose 在相同准确度度量方面, 却能分别提升 3.6%、1.3%、4.0%、3.2%、4.1%. 但本文方法较 HigherHRNet-W32 在网络参数量方面降低 8.7M、运算复杂度方面 GFLOPs 降低 44.95. 对比 Lite-HRNet 网络在 MS COCO 2017 验证集和测试集不同的实验结果对比, 由于本文方法的模型复杂程度更高, 在泛化能力方面表现较差, 所以存在验证集表现较优而测试集表现较差的情况. 但在实际运行效果中, L-MFNet 网络较其他方法能够直接运行在仅有 CPU 的低算力设备中, 帧率可达到 28 fps.

表 2 MS COCO 2017 测试集实验结果与其他方法对比

方法	Backbone	Input size	Params (M)	GFLOPs	mAP (%)	AP50 (%)	AP75 (%)	APM (%)	APL (%)
OpenPose[20]	—	—	—	—	61.8	84.9	67.5	57.1	68.2
Hourglass[7]	Hourglass	512	277.8	206.9	56.6	81.8	61.8	49.8	67.0
PersonLab[21]	ResNet-152	1401	68.7	405.5	66.5	88.0	72.6	62.4	72.3
Associative Embedding[22]	—	—	—	—	65.5	86.8	72.3	60.6	72.6
HigherHRNet-W32[8]	HRNet-w32	512	28.6	47.9	66.4	87.5	72.8	61.2	74.2
HigherHRNet-W48[8]	HRNet-w48	640	63.8	154.3	68.4	88.2	75.1	64.4	74.2
Lite-HRNet-18[19]	Lite-HRNet-18	384×288	1.1	0.45	67.6	87.8	75.0	64.5	73.7
Lite-HRNet-30[19]	Lite-HRNet-30	384×288	1.8	0.31	70.4	88.7	77.7	67.5	76.3
Ours	L-MFNet	512	19.9	2.95	65.4	86.2	71.5	60.3	72.3

3.5 消融实验

为了验证本文 L-MFNet 网络模块 (Shuffle-Neck、Ghost-Block、Attention module) 的有效性, 在 MS COCO 2017 训练集和验证集中, 进行进一步的验证分析实验, 结果如表 3 所示.

表 3 Shuffle-Neck 和 Ghost-Block 的消融实验

对比方法	Shuffle-Neck	Ghost-Block	Params (M)	GFLOPs	mAP (%)
HigherHRNet-W32	×	×	28.6	47.9	66.4
Shuffle-Neck	√	×	28.4	47.8	66.5
Ghost-Block	×	√	20.1	3.86	65.2
L-MFNet	√	√	19.9	2.95	65.4

从表 3 的消融实验数据分析: 仅改进 Shuffle-Neck 模块的 L-MFNet 网络, 由于 Shuffle-Neck 模块仅在 Stage1 阶段循环应用 4 次且在轻量化的同时加入 Attention 模块, 所以对网络轻量化的程度贡献较小, 仅有 0.2M 参数量和 0.1 GFLOPs 计算量. 但相较于 HigherHRNet 可以提升 0.2% mAP, 从侧面反映了双支路注意力特征融合模块的有效性. 使用 Ghost-Block 模块的 L-MFNet 网络仅以较低的准确率为代价, 得到了 7.5M 参数量和 44.04 计算量的降低. 通过引入 Ghost-Block 模块作为网络的基础, 利用不同特征层之间的冗余性将传统卷积层拆分成两部分, 按照一定比例一部分此案用传统卷基层, 另一部分部分则在之前基础上采用线性映射的计算得到剩余特征层. 实验结果证明, 引入 Ghost-Block 模块和双路注意力信息融合方法在以降低准确率 1.2% 为代价, 降低了网络 29.7% 参数量和 44.04 计算量的同时, 提升了模型的运行效率, 使网络能直接部署在低算力设备.

3.6 可视化分析

图 6(a) 分别为网络预测部分关键点热力图的真实值和预测值. 对比分析发现, 在参数量、计算量大幅降低情况下, 虽然存在预测准确度与真实值有所差异, 但仍可准确预测关键点的位置. 图 6(b) 是实际处理效果, 根据测试效果综合分析, 在多人场景下的关键点遮挡、截断、尺寸变化的情况下, 本文算法能够保持较强的关键点检测和分组能力. 但在图 1 部分关键点截断的情况下, 出现了关键点定位不准确的失误. 这说明了本文的轻量化方法, 在处理关节点截断方面还需进一步加强, 通过增加截断数据提升算法处理关键点截断的能力.

(a) 热力图可视化

(b) 检测效果可视化

图 6 网络预测结果可视化

4 结束语

本文以自底向上的姿态估计网络 HigherHRNet 为基础框架, 以 Shuffle-Neck 和 Ghost-Block 为基础, 构建了一种轻量化的多路特征融合网络. Shuffle-Neck 模块通过添加通道拆分和通道混洗构建轻量级残差模块, Ghost-Block 模块针对相同特征层之间存在的冗余性, 采用线性运算获取相同特征的不同特征层, Attention 模块采用平均池化和最大池化提取通道和空间注意力, 并提供了特征相加、通道拼接两种方式的注意力信息融合方法. 在 MS COCO 2017 测试集中, 相较于 HigherHRNet, 以 mAP 降低 1.0% 为代价, 将网络参数量方面降低 8.7M、运算复杂度方面 GFLOPs 降低 151.35. 如何保证轻量化高分辨率网络的同时, 进一步提高关键点检测的准确率是下一步研究的重点.

参考文献

1 Toshev A, Szegedy C. DeepPose: Human pose estimation via deep neural networks. Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014. 1653–1660.

2 Ahmed I, Jeon G, Chehri A, *et al*. Adapting Gaussian YOLOv3 with transfer learning for overhead view human detection in smart cities and societies. Sustainable Cities and Society, 2021, 70: 102908. [doi: 10.1016/j.scs.2021.102908]

3 Ren SQ, He KM, Girshick R, *et al*. Faster R-CNN: Towards

real-time object detection with region proposal networks. Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal: MIT Press, 2015. 91–99.

4 Lin TY, Dollár P, Girshick R, *et al*. Feature pyramid networks for object detection. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 936–944.

5 He KM, Gkioxari G, Dollár P, *et al*. Mask R-CNN. Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017. 2980–2988.

6 Sun K, Xiao B, Liu D, *et al*. Deep high-resolution representation learning for human pose estimation. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 5686–5696.

7 Newell A, Yang KY, Deng J. Stacked hourglass networks for human pose estimation. Proceedings of the 14th European Conference on Computer Vision. Amsterdam: Springer, 2016. 483–499.

8 Cheng BW, Xiao B, Wang JD, *et al*. HigherHRNet: Scale-aware representation learning for bottom-up human pose estimation. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 5385–5394.

9 Ma NN, Zhang XY, Zheng HT, *et al*. ShuffleNet V2: Practical guidelines for efficient CNN architecture design. Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich: Springer, 2018. 122–138.

10 Han K, Wang YH, Tian Q, *et al*. GhostNet: More features from cheap operations. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 1577–1586.

11 Zhao T, Wu XQ. Pyramid feature attention network for saliency detection. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 3080–3089.

12 Howard AG, Zhu ML, Chen B, *et al*. MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861, 2017.

13 Sandler M, Howard A, Zhu ML, *et al*. MobileNetV2: Inverted residuals and linear bottlenecks. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 4510–4520.

14 Howard A, Sandler M, Chen B, *et al*. Searching for MobileNetV3. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 1314–1324.

15 Zhang XY, Zhou XY, Lin MX, *et al*. ShuffleNet: An extremely efficient convolutional neural network for mobile devices. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 6848–6856.

16 Hu J, Shen L, Sun G. Squeeze-and-excitation networks. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7132–7141.

17 Fu J, Liu J, Tian HJ, *et al*. Dual attention network for scene segmentation. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 3141–3149.

18 Xiao B, Wu HP, Wei YC. Simple baselines for human pose estimation and tracking. Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich: Springer, 2018. 472–487.

19 Yu CQ, Xiao B, Gao CX, *et al*. Lite-HRNet: A lightweight high-resolution network. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 10435–10445.

20 Cao Z, Simon T, Wei SE, *et al*. Realtime multi-person 2D pose estimation using part affinity fields. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 1302–1310.

21 Papandreou G, Zhu T, Chen LC, *et al*. PersonLab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich: Springer, 2018. 282–299.

22 Newell A, Huang ZA, Deng J. Associative embedding: End-to-end learning for joint detection and grouping. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 2274–2284.

(校对责编: 孙君艳)