DOI:10.19651/j.cnki.emt.2415726

全局-局部特征融合的人体姿态估计算法*

毛 琳 任春贺 杨大伟
(大连民族大学机电工程学院 大连 116600)

摘 要:针对现有人体姿态估计算法存在因骨干网络特征提取不充分,导致关键点特征信息丢失的问题,提出一种结合全局-局部特征融合模块的人体姿态估计网络模型(GLF-Net)。为了在特征提取阶段获得高质量的特征图,该算法从全局特征和局部特征出发,对骨干网络 ResNet-50 进行改进,分别设计了全局极化自注意力模块和局部深度可分离卷积模块。同时采用并行的结构方式将融合了全局位置信息和局部语义信息特征的模块嵌入到骨干网络的 Bottleneck 层中,既能增强原骨干网络的特征提取能力,又为后续的 Transformer 网络提供有效的全局和局部特征输入,进而提高姿态关键点检测的性能。在公开人体姿态估计数据集 COCO 2017 上和 MPII 数据集上分别进行模型测试,该算法性能与与基准算法(Poseur)相比,姿态关键点的平均准确度(AP)提升了 2.1%,平均召回率(AR)提升了 1.5%,正确估计关键点比例(PCKh@0.5)最高达到 90.6。实验结果表明,所提算法在姿态估计精度上优于现存同类方法,可以明显提高人体姿态关键点的定位准确度。

关键词:人体姿态估计;特征提取;全局极化自注意力;局部深度可分离卷积;全局-局部特征融合
中图分类号:TN919 **文献标识码**:A **国家标准学科分类代码**:520.20

Global-local features fusion in human pose estimation algorithm

Mao Lin Ren Chunhe Yang Dawei
(College of Mechanical and Electronic Engineering, Dalian Minzu University, Dalian 116600, China)

Abstract: Aiming at the problem that the existing human pose estimation algorithm has insufficient feature extraction of the backbone network, which leads to the loss of key point feature information, a human pose estimation network model (GLF-Net) combined with global-local feature fusion module is proposed. In order to obtain high-quality feature maps in the feature extraction stage, the algorithm improves the backbone network ResNet-50 from the global and local features, and designs a global polarization self-attention module and a local depth separable convolution module respectively. At the same time, a parallel structure is used to embed the module that combines global position information and local semantic information features into the Bottleneck layer of the backbone network, which can not only enhance the feature extraction ability of the original backbone network, but also provide effective global and local feature input for the subsequent Transformer network, thereby improving the performance of pose key point detection. The model test is carried out on the public human pose estimation dataset COCO 2017 and MPII dataset respectively. Compared with the benchmark algorithm (Poseur), the average accuracy of the pose key points is increased by 2.1%, the average recall rate is increased by 1.5%, and the proportion of correctly estimated key points (PCKh@0.5) is up to 90.6. The experimental results show that the proposed algorithm is superior to the existing similar methods in the accuracy of pose estimation, and can significantly improve the positioning accuracy of human pose key points.

Keywords: human pose estimation; feature extraction; global polarized self-attention mechanism; local depth-wise separable convolution; global-local feature fusion

0 引 言

2D 人体姿态估计是指从给定的二维图像中估计人体关键点(如头部、手臂、腿部等关键节点),进而推断人体的姿态信息。现有的人体姿态估计算法大多采用普通卷积神经网络[1](convolutional neural network,CNN)和残差网

收稿日期:2024-03-28
* 基金项目:国家自然科学基金(61673084)、辽宁省自然科学基金(20170540192,20180550866,2020-MZLH-24)项目资助

络[2](residual network,ResNet)作为骨干网络用于特征提取。但对于复杂的姿态或受到遮挡、光照变化等情况,传统的 CNN 和 ResNet 会难以捕捉到某些细节特征,导致姿态估计的准确度较低。尽管近年来姿态估计算法取得了显著的进展,但仍然存在一些挑战和限制。

一方面,传统的基于卷积神经网络(CNN)的方法在提取局部特征方面表现出色,但往往忽略了全局特征的提取,导致在复杂场景下姿态估计的准确性下降。如 Xiao 等[3]提出的 Simple Baselines 采用了堆叠沙漏网络结构,并取消了跳跃连接以简化模型。Yu 等[4]、李一凡等[5]和梁桥康等[6]的研究都致力于通过优化网络结构来减少模型的参数量和计算复杂度,以实现更轻量级的姿态估计模型。但在提高计算效率的同时,却因简化网络结构而牺牲了对全局特征的处理能力,从而影响了关键点检测的准确性。Gong 等[7]、Cheng 等[8]和 Li 等[9]的研究方法都着重于通过多尺度特征融合来提高姿态估计的性能,但由于对全局特征的忽略限制了其在复杂场景下的性能。石跃洋等[10]的方法则通过添加尺度转换结构来优化多尺度检测能力,但同样受限于全局特征的提取能力,可能导致在关键点检测过程中出现误检或漏检的问题。此外,Ma 等[11]提出的 PPT 方法虽然通过融合多视图特征来增强特征表达,但这种方法也未能充分考虑全局姿态的上下文信息,从而影响了姿态估计的准确性。Ren[12]在 PPT[11]基础上提出的 DPPT 方法利用自蒸馏技术提高了姿态估计效率,但在全局和局部特征的综合考虑上仍有待加强。

另一方面,虽然注意力机制[13]能够捕捉全局信息,但单独使用注意力机制可能会增加计算复杂性,并且忽略了局部细节信息。如 Hua 等[14]和 Bao 等[15]分别采用残差注意力机制和双重注意力机制来增强网络的特征表示,通过给有用信息分配更高的权重来减少无关信息的干扰。此外,Ou 等[16]提出添加全局上下文模块对编码器和解码器进行全局增强,而 Zhang 等[17]直接在 HRNet 中加入注意力机制来增强网络的特征表示,使输出更准确地估计人体关节点的位置。方等[18]提出在编码层中加入通道注意力机制,解码层则采用多个反卷积模块,以在降低模型复杂度的同时保持预测精度。Ye 等[19]提出通过注意力矩阵将学生模型和教师模型之间的视觉标记和关键点标记对齐,实现了学生模型在学习教师模型知识的同时保持结构优势。尽管这些方法都取得了一定的效果,但注意力机制的一个主要缺点是它需要对每个位置进行单独计算,这大大增加了计算的复杂性。因此为了避免这种情况,一些研究者提出使用自注意力机制[20]替代它,该机制可以同时计算输入序列中所有位置之间的关系,计算效率更高。Sun 等[21]、Li 等[22]、Yuan 等[23]和 Xu 等[24]均采用 Transformer 模型,并充分利用自注意力机制的特性,来学习关注不同输入特征中人体特定的区域。此外,Li 等[25]和 Zhou 等[26]分别在 Mask-RCNN 和 HRNet 模型中引入自注意力机制,用于提取全局信息。尽管这证明了自注意力机制在特征提取方面的优势,但仍忽略了图像中人体的整体特征与局部特征信息之间的互补相关性。这种忽略可能导致姿态估计结果中细节语义信息的不足,从而影响了姿态估计的准确性。

为了解决这两种问题,学者们开始探索如何更好地结合全局特征和局部特征的优势,目前已有部分研究提出这种特征补足的思想,旨在不同层次的信息间构建联系。Chen 等[27]提出一个动态的卷积算子 X-volution,通过多分支拓扑结构将由卷积和自注意力组成的多分支结构进行融合;Pan 等[28]提出了一种混合算子 ACMix,通过共享相同的重操作来整合自注意和卷积模块。目前这两种方法仅在图像分类、目标检测和语义分割领域上进行过实验。Zhao 等[29]提出的方法 DPIT,通过自底向上的分支和自上而下的分支分别捕获局部结构信息和全局依赖关系,并利用 Transformer 的编码器来探索这种交互作用。现阶段,类似上述的融合方法在人体姿态估计领域的应用相对较少,且因其复杂的网络架构和算法设计,大部分网络模型对显卡设备的性能提出了较高要求,因此这些方法通常需要使用配备高性能 GPU 的计算机来支持其计算密集型的操作,以确保在训练和推理过程中的高效性和准确性。而在普通设备上运行模型很可能会导致处理速度极慢、内存不足甚至无法正常运行的情况,从而显著影响姿态估计的性能。

基于此,为解决因人体关键点特征提取不充分而导致关键点定位不准确的问题,本文在 Mao 等[30]的基准上提出了基于全局-局部特征融合网络(global-local feature fusion network,GLF-Net),通过优化网络中模块的性能以达到充分提取特征的目的。具体而言,本文提出的全局极化自注意力机制(global polarized self-attention,G-PSA)和局部深度可分离卷积模块(local depth-wise separable convolution,L-DSC)分别优化了全局和局部关键点特征的提取,还通过全局与局部特征互补的提取模块(convolution and self-attention fusion module,CSAF)有效地融合了这两种特征。这种全局与局部特征的平衡策略,使得本文的方法在保持一定计算复杂性的同时,能够更准确地定位人体关键点,提高了姿态估计的性能。与同类增强方法相比,本文提出的算法与之显著的区别在于它充分考虑了骨干网络对全局和局部特征的提取,并采用了合适的融合策略来实现特征的增强和互补。此外,该方法在资源受限的普通设备上同样能够保持出色的性能,具有更广泛的应用前景。

1　GLF-Net 算法

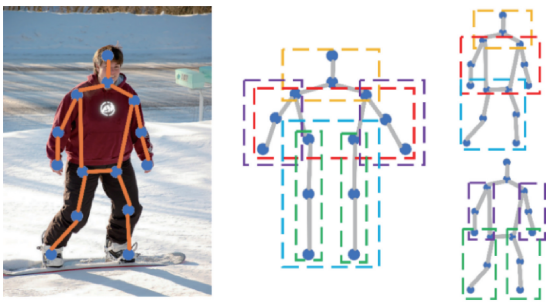
1.1　问题分析

关键点特征提取是指从人体图像中获取关键点的位置信息,用于描述人体的姿态。这些关键点通常是人体的关节、身体部位或其他重要的特征点,例如头部、脚踝等。人体姿态估计的基础是关键点特征提取,因此只有准确地提取出关键点的位置信息,才能进一步计算出人体的姿态。

在传统的姿态估计网络中,CNN 通常使用多个卷积层和池化层进行堆叠,以提取图像的局部特征,但是它的局部特征提取能力有限,无法提取出更全局、更抽象的特征。并且卷积核之间可能会有一些相似的特征提取能力,导致它们在处理图像时会产生类似的输出,容易产生冗余,还会影响模型的性能和效率,这就需要使用其他方法来提取更全局、更抽象的特征,以提高姿态关键点检测的能力。因此,对于 2D 人体姿态估计任务来说,针对特征提取这一问题进行深入研究具有重大意义。

特征提取可以分为对全局信息和局部信息的提取。全局特征(global features)是指对每张人体图片的整体信息进行特征提取,其中不包含任何空间信息,只包含整张图片的统计信息,如颜色、纹理和形状等。然而,当人体姿态没有正确对齐时,会导致关键信息在图像中出现偏移或错位的情况,从而使全局特征无法准确表达人体姿态信息。为了解决这个问题,提出了局部特征(local features)。局部特征是指从图像的某个区域中提取出的特征,比如人体的头部、上半身、下半身等局部区域的特征。通过提取多个局部特征并进行融合,可以获取更全面、更准确的信息。

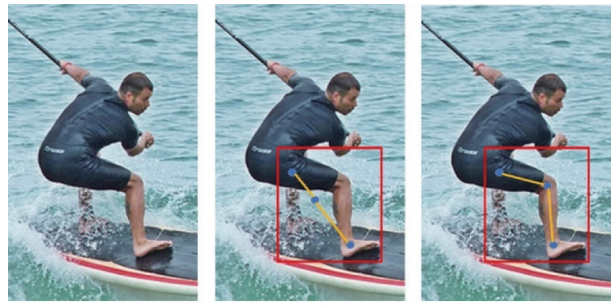
以标注 14 个关键点为例,如图 1 所示为人体姿态估计中的全局和局部关键点特征表示。这些关键点被用来提取人体结构中的 7 个 ROI(region of interest),分别对应图 1(b)中的头、上半身、下半身、左臂、右臂、左腿和右腿。使用同一个 CNN 对 ROI 区域和原始图片分别提取关键点信息,可以得到如图 1(a)所示的全局特征和如图 1(b)所示的七个局部特征。



(a) 全局特征　　(b) 局部特征

图 1　全局和局部关键点特征表示

在人体姿态估计任务中,如果全局特征提取不足,会导致部分关键点无法被正确定位和识别。若算法只考虑了局部特征,例如只关注腿的弯曲程度,而没有考虑到身体的整体姿态和朝向,如图 2(b)的矩形框所示,那么算法可能会将弯曲的手臂识别为伸直的手臂,从而导致姿态估计不准确。然而人体姿态又具有多变性,有些关键点可能会被遮挡。若只考虑了全局特征,对局部特征提取不足,可能会导致无法准确检测和定位被遮挡的部位。如图 3(b)的矩形框表示,由于算法没有对肘关节进行充分的特征提取,导致肘部的位置无法准确地被检测出来,从而影响整个姿态的估计结果。



(a) 原图　　(b) 错误检测图　　(c) 正确检测图

图 2　全局特征提取不足的示例图

(a) 原图　　(b) 错误检测图　　(c) 正确检测图

图 3　局部特征提取不足的示例图

在 2D 人体姿态估计任务中,全局特征可以提供人体的整体轮廓信息、姿态角度等高层次特征,而局部特征则可以提供更加微观的关节位置、角度等低层次特征。因此为解决现有算法存在因骨干网络特征提取不充分而导致关键点定位不准确的问题,本文在特征提取方面同时考虑全局特征和局部特征,提出改进后的全局极化自注意力机制和局部深度可分离卷积,通过多分支结构同时提取全局特征和局部特征,并使用简单的加法融合方式,即将两个特征向量的每个元素相加,得到一个新的特征向量作为融合结果,以更好地捕捉特征信息。通过将不同层次、不同范围的特征进行整合,可生成细节更丰富的图像。该方法旨在实现对人体的各个关节进行检测和准确定位,提高人体姿态模型对关键点的理解和识别能力,进而提升姿态估计的准确性。

1.2　全局极化自注意力

2D 人体姿态估计的目标是通过分析一张图像预测并定位出人体各个关键点的位置。因此需要对每个关键点像素进行回归。而利用 ResNet-50 作为姿态估计骨干网络所提取的特征在像素级回归上不够充分,所以如何得到更全局的像素级特征表示至关重要。由于自注意力机制模块能够捕获通道信息和位置信息之间的内在联系,从而使得网络能够更好地区分不同特征的信息,因此本文算法借鉴极化自注意力机制[31]在提取全局特征时充分考虑像素的回归的思想,在该模块的基础上进行结构改进,构成了全局极化自注意力(G-PSA)模块,如图 4 所示。

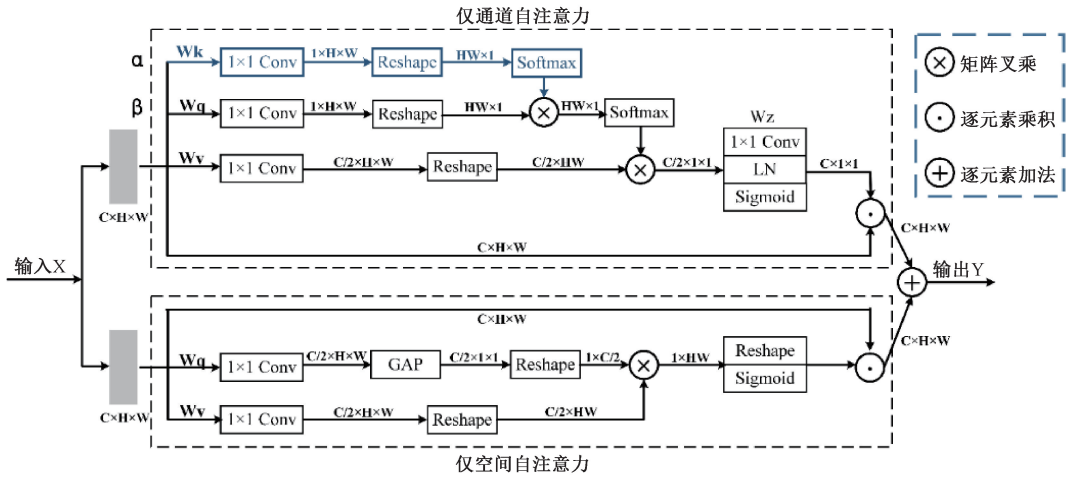


图 4　全局极化自注意力结构

G-PSA 具体包括两个部分:仅通道自注意力和仅空间自注意力。本文是在仅通道极化自注意力分支中进行改进,如图 5 所示,增加了一条卷积支路 α,这条支路可被称为“修正支路”。它是通过引入额外的信息学习得到最佳的权重,以便在注意力计算中对支路 β 原有的权重进行修正,修正后的输出可以更好地反映人体姿态的结构特征,以帮助模型更好地理解人体姿态,从而提高姿态估计的准确性。

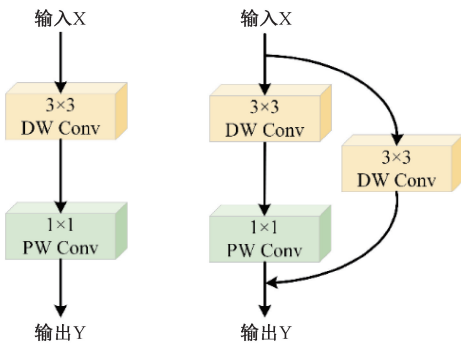


图 5　局部深度可分离卷积结构

G-PSA 模块是将输入的特征图分成两个子集,分别在通道和空间上进行自注意力计算,然后将两个子集的结果进行加权求和得到最终的自注意力表示。这种方法可以有效地提高自注意力的表达能力,特别是在处理像素级别的任务时,可以更好地捕捉到全局细节信息,从而提高模型的性能。

仅通道极化自注意力分支具体可以表达为如下公式:

$$C_k = F_{sm}(R_1(W_k(X))) \tag{1}$$

$$C_q = F_{sm}(R_2(W_q(X)) \otimes C_k) \tag{2}$$

$$C_v = R_3(W_v(X)) \otimes C_q \tag{3}$$

$$A^{ch} = F_{sg}(LN(W_z(C_v))) \tag{4}$$

$$Z^{ch} = A^{ch} \odot^{ch} X \tag{5}$$

其中,给定一个输入 $X \in \Re^{C\times H\times W}$,$W_k$、$W_q$、$W_v$ 和 W_z 分别为标准的 1×1 卷积,R_1、R_2 和 R_3 分别表示三种维度整形算子,F_{sm} 和 F_{sg} 分别代表 Softmax 运算和 Sigmoid 运算,LN 是层归一化运算,$C_k \in \Re^{HW\times 1}$、$C_q \in \Re^{HW\times 1}$ 和 $C_v \in \Re^{\frac{C}{2}\times 1\times 1}$ 分别是三条支路经过各种运算后的输出,$A^{ch} \in \Re^{C\times 1\times 1}$ 是仅通道分支的极化自注意力运算,$Z^{ch} \in \Re^{C\times H\times W}$ 为仅通道分支的输出,$\otimes$ 和 $\odot^{ch}$ 分别为矩阵叉乘和通道逐元素乘积运算。

仅空间极化自注意力分支具体可以表达为如下公式:

$$S_q = R_1(F_{gap}(W_q(X))) \tag{6}$$

$$S_v = R_2(W_v(X)) \tag{7}$$

$$A^{sp} = F_{sg}(R_3(S_q \otimes S_v)) \tag{8}$$

$$Z^{sp} = A^{sp} \odot^{sp} X \tag{9}$$

其中,给定一个输入 $X \in \Re^{C\times H\times W}$,$W_q$ 和 W_v 分别为标准的 1×1 卷积,R_1、R_2 和 R_3 分别表示三种维度整形算子,F_{gap} 和 F_{sg} 分别代表自适应平均池化操作和 Sigmoid 运算,$S_q \in \Re^{1\times\frac{C}{2}}$ 和 $S_v \in \Re^{\frac{C}{2}\times HW}$ 分别是两条支路经过各种运算后的输出,$A^{sp} \in \Re^{1\times H\times W}$ 是仅空间分支的极化自注意力运算,$Z^{sp} \in \Re^{C\times H\times W}$ 为仅空间分支的输出,$\otimes$ 和 $\odot^{sp}$ 分别为矩阵叉乘和空间逐元素乘积运算。

因此,经过全局极化自注意力(G-PSA)模块将仅通道自注意力和仅空间自注意力进行⊕加权求和得到最终输出 $Y \in \Re^{C\times H\times W}$,表达为如下公式:

$$Y = Z^{ch} \oplus Z^{sp} \tag{10}$$

在 2D 人体姿态估计任务中,改进后的 G-PSA 模块可以根据特征图中每个位置的重要性,自适应地调整不同位置的权重,从而更好地捕捉图像中的关键特征,帮助模型更好地理解图像中的全局语义信息。与传统的注意力机制相比,极化自注意力模块能够更好地处理长距离依赖关系,同时还能够减少计算量,提高模型的效率。

1.3　局部深度可分离卷积

本文算法借鉴深度可分离卷积[32]在提取局部特征时

能保持模型轻量化的同时,还能提高模型感受野和局部特征提取能力的思想,对该卷积模块进行结构改进,并结合残差思想,构成新的局部深度可分离卷积(L-DSC)结构,如图 5(b)所示,以加强深度可分离卷积的特征提取能力。

具体来说,为了提取更加细致的局部特征,本文在原有模块基础上增加了一个带残差连接的 3×3 卷积模块,旨在整合不同通道在相同位置上的特征信息,在保留原始局部特征信息的同时,增强深度可分离卷积在通道之间的信息融合能力,以便提取丰富的局部信息,有利于更好地捕捉人体姿态中的细节特征。

L-DSC 模块具体可以表达为如下公式:

$$Y_1 = F_{PW}(F_{DW}(X)) \tag{11}$$

$$Y_2 = F_{PW}(F_{DW}(X)) + F_{DW}(X) \tag{12}$$

其中,给定一个输入 $X \in \Re^{C\times H\times W}$,$F_{DW}$ 为 3×3 的深度卷积,F_{PW} 为 1×1 的逐点卷积,$Y_1 \in \Re^{C\times H\times W}$ 和 $Y_2 \in \Re^{C\times H\times W}$ 分别为改进前和改进后的最终输出。

在 2D 人体姿态估计任务中,改进后的 L-DSC 模块可以保持特征信息输入的完整性,确保传递到下一阶段的特征仍然包含重要信息,避免信息丢失,增添的 3×3 卷积有利于提取更丰富的局部关键点特征。该模块可以在不增加网络参数量的同时增强模型的特征提取能力,利用丰富的局部特征更好地捕捉通道之间的相关性,进而提高姿态估计的准确性。

1.4　全局-局部特征融合

对于原骨干网络 ResNet-50 来说,特征提取路径形成了细节信息丰富的全局特征和语义信息丰富的局部特征。本模块设计的目的是为了将二者的优势相结合,减少特征在传输过程中的信息损失,提高骨干网络的特征表达能力和非线性拟合能力。因此,为了更好地捕捉人体姿态的细节和整体结构,本文考虑借鉴融合卷积和自注意力机制的方式,使得骨干网络能够同时考虑局部和全局信息,以便更准确地预测人体姿态。

如图 6(a)所示为 ResNet-50 的原始残差块,本文算法是在其 3×3 卷积之后加入改进后的全局极化自注意力(G-PSA)和局部深度可分离卷积(L-DSC),使二者构成如图 6(b)所示的高效信息融合模块(CSAF),用于融合全局和局部特征信息,以增强网络对姿态特征的捕捉能力。

本文算法的模型结构主要分为两部分,一部分是全局极化自注意力(G-PSA)机制,其改进策略是在仅通道自注意力的分支上增加一条修正支路,为提取的上下文全局特征表示分配更多的权重,通过对每个像素点的特征进行加权来加强重要像素点的特征,从而更好地捕捉全局上下文信息,理解图像中的姿态变化;另一部分是局部深度可分离卷积(L-DSC)模块,其改进方法是在原来的基础上增加一个带残差连接的卷积层支路,利用不同大小的卷积核来进一步提取局部特征,以增强对细节信息的感知能力。然后,把经过两个模块分别得到的特征向量进行特征级信息融合。最后,将融合后的特征经过进一步地有效处理,送入到后续的网络中来预测人体姿态的关键点。这种融合方式可以在保留局部特征的同时,充分利用全局信息,从而实现对人体姿态的准确估计。

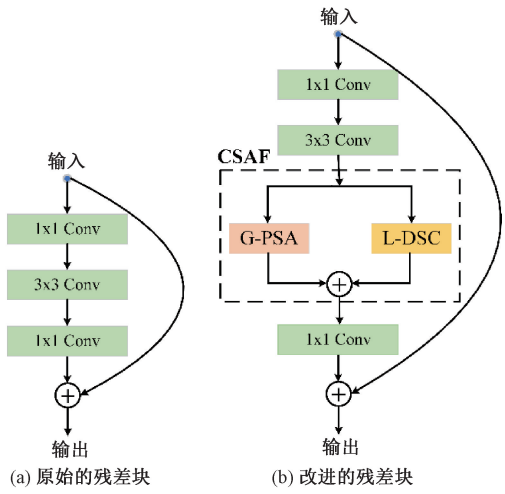


图 6　骨干网络残差块结构

由于深度残差网络 ResNet-50 中的深层语义信息在特征提取过程中不断地被传递,深层语义信息在浅层网络中起到的作用逐渐减弱。因此,本文算法将特征融合模块(CSAF)嵌入至 ResNet-50 的每一个残差块中作为新的骨干网络,使模型同时保留深层的局部特征语义信息和低层的全局特征细节信息,并加强有用通道的权重,以确保尽可能地减少特征信息的损失,从而生成更加丰富的特征表示。

1.5　整体网络

本文的整体网络架构主要包括用于特征提取的 CNN 网络和特征后处理的 Transformer 网络两个部分。特征提取部分是由全局-局部特征融合网络 GLF-Net 构成;而特征后处理部分是由关键点编码器和查询解码器组成的 Transformer 网络构成,模型总体架构如图 7 所示。

在人体姿态估计网络的特征提取部分,首先将一张 RGB 图像输入到 GLF-Net 中以提取丰富的语义特征信息,通过建立深层特征通道之间的非线性关系和空间位置关联信息,聚焦于重要区域和目标特征。本文算法采用 Resnet-50 作为骨干网络主要由四个残差块组成,在图 7 中依次简称为 ResBlock1、ResBlock2、ResBlock3 和 ResBlock4。其次,通过特征金字塔自顶而下的融合方式在每个残差块都结合 CSAF 模块,该模块是将提供整体姿态信息的全局特征和提供更细节信息的局部特征进行融合,以便更好地捕捉人体姿态的细节和整体特征,形成更丰富的特征表示。最后,将融合后的特征信息输入到 Transformer 中。在关键点特征的后处理部分,主要是利用骨干网络提取的特征信息获得密集的特征映射来预测关键点坐标,使得 Transformer 获得融合全局和局部特征

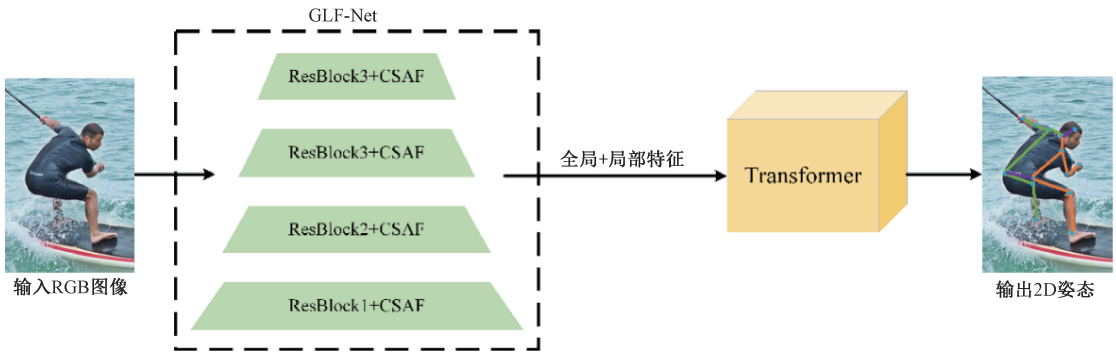


图 7　整体网络框架

的有效输入，从而提高姿态估计网络的准确性。

2　实验结果分析

2.1　数据集

针对 2D 人体姿态估计任务，实验采用 COCO 2017[33] 和 MPII[34] 关键点检测数据集中的 RGB 图像进行训练和测试。COCO 2017 包含大约 80 类图像和大约 250 000 个人物实例，该数据集由图像文件和注释文件组成，一共定义了 17 个人体关键点，标注信息包括人体关键点的位置和可见性。在仿真中使用 COCO 训练集中的 118 287 张图像进行训练，验证集中的 5 000 张图像进行验证和测试。MPII 包含超过 25 000 张人类活动的各种姿态图像，每张图像都定义了 16 个人体关键点，包括头部、手臂、腿部等身体部位的关键点。

2.2　实验设计

本文操作系统为 Ubuntu 20.04，深度学习框架为 Pytorch 1.11.0、Python 3.8 和 CUDA 11.3，使用 6 张 NVIDIA GeForce RTX 4090 显卡用于训练和测试人体姿态估计模型。在训练网络模型时，将批尺寸设置为 24，采用 AdamW 优化器，初始学习率为 7.5×10^{-4}，衰减权重为 7.5×10^{-5}。对于 COCO 和 MPII 数据集，在训练和测试时将数据集中的图像尺寸均设置为 256×192，设置训练迭代次数为 325 次。在训练过程中，每经过 25 个 epoch 就会保存一次模型权重，最终通过综合比较这些训练模型的精度和推理速度，从中选出最优模型用于测试以得出最后的衡量指标。

为评估算法表现，体现 2D 人体姿态估计算法对关键点的检测效果，COCO 数据集主要采用平均准确率 AP (average precision)值和平均召回率 AR(average recall)作为衡量标准检验姿态估计的精度。在姿态估计中，AP 和 AR 通常是通过计算预测关键点与真实关键点之间的距离来计算的，即以关键点相似性(object keypoint similarity，OKS)，如果距离小于一定阈值，则认为该关键点被正确检测，否则认为该关键点未被检测。OKS_p 可表示为：

$$OKS_p=\frac{\sum_i \exp\{-d_{p_i}^2/2S_p^2\sigma_i^2\}\delta(v_{p_i}=1)}{\sum_i \delta(v_{p_i})} \tag{13}$$

其中，p 为真值人的 ID；i 为关键点的 ID；d_{p_i} 为每个人的真实关键点位置与预测关键点位置的欧氏距离；S_p 为当前真值人所占面积的平方根；σ_i 为第 i 个关键点的归一化因子；v_{p_i} 为第 p 个人的第 i 个关键点是否可见；δ 为克罗内克函数。

MPII 所采用的评价标准是 PCKh (percentage of correct keypoint of the head)，用于衡量关键点的正确估计比例。它使用的它是以当前人的头部直径作为尺度因子，即头部左上点与右下点的欧式距离，在一定阈值下计算关键点正确率。一般情况下，选择人体头部半径的 0.5 倍作为阈值，即 PCKh@0.5。如果一个关键点的估计距离真实位置小于阈值，则认为该关键点被正确估计。该指标的取值范围为 0 到 1，值越大表示算法的检测效果越好。PCK_i^k 用公式可表示为：

$$PCK_i^k=\frac{\sum_p \delta(\frac{d_{pi}}{d_p^{def}}\leqslant T_k)}{\sum_p 1} \tag{14}$$

其中，p 表示第 p 个人，T_k 表示人工设定的阈值，k 表示第 k 个阈值，d_{pi} 表示第 p 个人的第 i 个关键点预测值与人工标注值之间的欧式距离，d_p^{def} 表示第 p 个人的尺度因子，δ 为克罗内克函数。

2.3　仿真结果与分析

为证明模型的优越性，本文将改进网络 GLF-Net 分别在 COCO 2017 和 MPII 数据集上进行了对比，测试结果分别如表 1 和表 2 所示。表 1 选取公认的基准模型分别为 SimpleBaseline[3]、Lite-HRNet[4]、AutoPose[7]、Dite-HRNet[9]、ESBN[15]、DistilPose[19]、HRFormer[23]、X-HRNET[26] 和 DPIT[29]，主要是为了全面评估所提出的 GLF-Net 模型在人体姿态估计任务上的优越性和性能。选取的算法涵盖了从传统的卷积神经网络(CNN)到基于 Transformer 的模型，以及轻量级和高效的网络结构，这些算法在人体姿态估计领域都具有一定的代表性。

在 256×192 大小的输入下，GLF-Net 引入了自注意力和卷积模块的结合，能够捕获更丰富的全局和局部特征。其总体指标精度基本优于其他模型方法，且 AP 值最

表 1　COCO 2017 数据集在不同方法下的对比结果

模型	Backbone	Params	GFLOPs	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
SimpleBase[3]	ResNet-50	34.0 M	8.90 G	70.4	88.6	78.3	67.1	77.2	76.3
SimpleBase[3]	ResNet-101	53.0 M	12.40 G	71.4	89.3	79.3	68.1	78.1	77.1
SimpleBase[3]	ResNet-152	68.6 M	15.70 G	72.0	89.3	79.8	68.7	78.9	77.8
Lite-HRNet[4]	Lite-HRNet-18	1.1 M	205.2 M	64.8	86.7	73.0	62.1	70.5	71.2
Lite-HRNet[4]	Lite-HRNet-30	1.8 M	319.2 M	67.2	88.0	75.0	64.3	73.1	73.3
AutoPose[7]	AutoPose	—	10.65 G	73.6	90.6	80.1	69.8	79.7	78.1
Dite-HRNet[9]	Dite-HRNet-18	1.1 M	0.2 G	65.9	87.3	74.0	63.2	71.6	72.1
Dite-HRNet[9]	Dite-HRNet-30	1.8 M	0.3 G	68.3	88.2	76.2	65.5	74.1	74.2
ESBN[15]	ResNet-34	33 M	21.3 G	72.0	91.1	79.9	68.9	77.7	77.5
ESBN[15]	ResNet-50	—	—	72.3	91.3	80.1	69.1	78.2	77.8
ESBN[15]	ResNet-101	—	—	72.6	91.6	80.5	69.4	78.4	78.1
ESBN[15]	ResNet-152	—	—	73.3	91.3	80.8	70.1	78.9	78.8
DistilPose-S[19]	stemnet	5.4 M	2.38 G	71.0	91.0	78.9	67.5	76.8	—
DistilPose-L[19]	HRNet-W48-stage3	21.3 M	10.33 G	73.7	91.6	81.1	70.2	79.6	—
HRFormer[23]	HRFormer-T	2.5 M	1.3 G	70.9	89.0	78.4	67.2	77.8	76.6
HRFormer[23]	HRFormer-S	7.8 M	2.8 G	74.0	90.2	81.2	70.4	80.7	79.4
X-HRNET[26]	X-HRNET-18	1.3 M	194.4 M	65.1	86.7	72.7	62.3	70.9	71.2
X-HRNET[26]	X-HRNET-30	2.1 M	300.2 M	67.4	87.5	75.4	64.5	73.3	73.5
DPIT[29]	DPIT-B	20.8 M	—	73.6	91.4	81.2	70.4	79.5	78.9
DPIT[29]	DPIT-L	38.0 M	—	74.6	**91.9**	82.1	71.3	80.6	79.9
Poseur[30]	ResNet-50	33.26 M	4.52 G	73.2	89.2	79.8	69.5	79.9	78.9
GLF-Net(本文)	ResNet-50	50.94 M	5.06 G	75.3	90.9	**82.3**	**71.7**	**81.7**	**80.4**

表 2　MPII 数据集在不同方法下的对比结果(PCKh@0.5)

模型	Backbone	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Mean
SimpleBase[3]	ResNet-50	96.4	95.3	89.0	83.2	88.4	84.0	79.6	88.5
SimpleBase[3]	ResNet-101	96.9	95.9	89.5	84.4	88.4	84.5	80.7	89.1
AutoPose[7]	AutoPose	96.6	95.0	88.3	83.2	87.2	82.8	78.9	88.0
PPT[11]	PPT-S	96.6	94.9	87.6	81.3	87.1	82.4	76.7	87.3
PPT[11]	PPT-B	97.0	95.7	90.1	85.7	89.4	85.8	81.2	89.8
DPPT[12]	DPPT-S	96.4	94.9	88.3	81.8	88.2	83.0	78.3	87.9
文献[18]	—	96.4	95.2	89.2	84.3	88.8	84.8	80.4	89.0
TokenPose[22]	TokenPose-S	96.0	94.5	86.5	79.7	86.7	80.1	75.2	86.2
TokenPose[22]	TokenPose-B	97.0	96.1	90.1	85.6	89.2	86.1	80.3	89.7
TokenPose[22]	TokenPose-L/D6	97.1	95.9	91.0	85.8	89.5	86.1	82.7	90.1
TokenPose[22]	TokenPose-L/D12	**97.2**	95.8	90.7	**85.9**	89.2	86.2	82.3	90.1
TokenPose[22]	TokenPose-L/D24	97.1	95.9	90.4	86.0	89.3	87.1	82.5	90.2
DPIT-B[29]	DPIT-B	97.1	95.7	90.0	84.6	89.4	85.9	80.7	89.6
DPIT-L/D6[29]	DPIT-L	96.7	95.9	90.8	**85.9**	89.2	86.0	82.6	90.1
Poseur[30]	ResNet-50	96.3	96.0	90.8	85.7	89.7	86.7	82.0	90.1
GLF-Net(本文)	ResNet-50	96.6	**96.4**	**91.1**	85.8	**90.3**	**87.4**	**82.8**	**90.6**

高达到 75.3%,AR 值最高达到 80.4%,取得了较好的准确度。SimpleBaseline[3]是一个基础的姿态估计模型,其简单而高效的结构为许多后续研究提供了参考。该模型的主干网络分别应用了 ResNet-50、ResNet-101 和 ResNet-152 三种不同的网络结构,随着残差网络层数的增加,更深层的网络结构往往具有更多的参数量和计算量,在这种情况下,使用更大的网络并不能带来更好的性能提升;与 Lite-HRNet[4]、Dite-HRNet[9]和 X-HRNET[26]相比,这些

模型都是基于 HRNet(高分辨率网络)的轻量级变体,旨在提高模型效率同时保持高分辨率表示。尽管 GLF-Net 不如它们轻量级,但本文算法是通过引入新的融合模块和更优化高效的结构,可以更好地捕捉更丰富的图像特征,实现模型在姿态估计领域的性能;与 AutoPose[7] 相比,该方法是采用神经网络搜索技术来优化模型结构,这与 GLF-Net 追求高效模型设计的目标相一致。虽然 GLF-Net 的参数量和计算量略高,但它的总体性能在 AP 和 AR 两种主要指标下均有较好的改善。而 AutoPose[7] 由于搜索空间的增大,可能更容易增加超参数的数量和模型的复杂性,同时过度依赖搜索也可能导致资源的浪费;与 ESBN[15] 相比,该方法是一种编码器-解码器网络,并关注多尺度特征的融合,这与本文提出的全局和局部特征融合有一定的相似性。该模型主要依赖于空间注意力模块,而 GLF-Net 结合了自注意力和卷积,能够更全面地捕获人体的重要特征信息;与 DistilPose[19] 相比,该方法通过知识蒸馏技术来提高人体姿态估计的性能,这也是一种有效的模型优化策略。然而 GLF-Net 采用了不同的模块优化策略,通过引入自注意力和卷积的融合模块,直接优化模型结构,同时性能结果比 DistilPose[19] 更好;与 HRFormer[23] 相比,该模型结合了高分辨率网络和 Transformer。HRFormer[23] 是以窗口自注意力的 Transformer 为主模块,用于高分辨网络的深层特征提取,而 GLF-Net 在主干网络中融合了自注意力和卷积模块,为后续的 Transformer 编码器提供了更丰富的输入特征,并且其解码器采用了交叉注意力机制,进一步增强了特征之间的融合交互,本文方法在不同的指标下均有较好的表现;与 DPIT[29] 相比,该方法同样是一个结合 Transformer 的模型,在框架中实现了自下而上分支和自上而下分支的有效融合。然而由于 GLF-Net 引入了本文提出的全局极化自注意力和局部深度可分离卷积的融合机制,使其在处理关键点之间的复杂关系时更加有效,模型的表现性能更为优越。

表 2 为 GLF-Net 与其他人体姿态估计方法在 MPII 测试数据集上在 PCKh@0.5 的标准下的性能比较,选取公认的基准模型分别为 SimpleBaseline[3]、AutoPose[7]、PPT[11]、DPPT[12]、文献[18]、TokenPose[22] 和 DPIT[29]。在输入尺寸同样为 256×192 的情况下,尽管 GLF-Net 方法在参数量上均略高于其他对比先进的方法,但本文算法在指标关键点平均准确度上最高达到 90.6,比原基准模型的指标 PCKh 值提高了 0.5。与 SimpleBaseline[3]、AutoPose[7]、PPT[11]、DPPT[12] 和文献[18]模型相比,尽管这些模型采用了不同的主干网络进行训练和测试,但通过在复杂人体姿态上的性能结果来看,本文模型在不同身体部位的指标精度明显优于这些模型。这反映出本文模型在处理复杂姿态时能够更好地捕捉关键点的位置和结构信息,从而提高了姿态估计的准确性。而与 TokenPose[22] 和 DPIT[29] 相比,尽管在头部和手腕的得分略有差距,但是 GLF-Net 在其他部位的检测得分均比其他架构的方法表现性能都要好。由此表明,本文所提算法在各个身体部位的检测效果较好,有效提高了人体姿态估计关键点检测的准确度,证明了算法的先进性。

2.4　消融实验

在人体姿态估计领域,有效融合全局特征和局部特征对于提高模型性能至关重要。本文通过在 COCO 2017 数据集上进行消融实验,来验证全局极化自注意力与局部深度可分离卷积的融合模块(CSAF)在特征提取上的有效性。首先,本文选取了 Poseur 作为基准算法,通过引入不同的模块进行对比实验。具体而言,C-Poseur 方法只添加了局部深度可分离卷积模块,SA-Poseur 方法则只引入了全局极化自注意力模块,而 GLF-Net 方法则是在模型中同时添加了全局极化自注意力和局部深度可分离卷积的融合模块。通过比较这些模型的性能,可以深入了解全局特征与局部特征对人体姿态估计的影响。

表 3 中展示了 3 种模型在 COCO 2017 数据集上的实验结果,其中列出了关键的性能指标,包括平均准确率(AP)和平均召回率(AR),同时还提供了模型的算力和参数量信息。在 256×192 的输入尺寸下,C-Poseur 和 SA-Poseur 相较于原始 Poseur 方法均取得了一定的检测效果提升,C-Poseur 方法的 AP 提高了 1.6%,AR 提高了 1.0%,SA-Poseur 方法的 AP 提高了 1.9%,AR 提高了 1.4%。而引入全局-局部特征融合模块(CSAF)的 GLF-Net 方法,取得了更为显著的人体姿态估计效果,该方法的 AP 提高了 2.1%,AR 提高了 1.5%。

表 3　基于模块选择的性能研究在 COCO 2017 数据集上的消融实验

方法	+G-PSA	+L-DSC	Params	GFLOPs	AP	AR
Poseur	—	—	33.26 M	4.52 G	73.2	78.9
C-Poseur	—	√	34.56 M	4.75 G	74.8	79.9
SA-Poseur	√	—	49.64 M	4.83 G	75.1	80.3
GLF-Net	√	√	50.94 M	5.06 G	**75.3**	**80.4**

为了更好的设计融合模块,本文在表 4 中进行了对比实验。从结果上可以看出,3 种不同的融合策略虽然在一定程度上都能提升原始模型的性能,但相加融合方法产生的效果最优。这是因为全局特征通常捕获了整幅人体图像或较大区域的信息,而局部特征则更关注于特定区域或关键点的细节,本文采用的相加融合方式更直接地结合了全局和局部特征,保留了人体关键点的位置信息,使得模型更容易学习到有效的表示。而连接融合方法是将全局特征和局部特征在通道维度上进行连接,从而形成一个更宽的特征图。这种方法能够虽然保留了所有特征的信息,但会增加特征图的通道数,从而使模型训练更加复杂且难

以优化;相乘融合则会改变特征图中每个位置的值,这可能会导致某些特征信息丢失。

表 4　模块的不同融合方式在 COCO 2017 数据集上的消融实验

方法	Params	GFLOPs	AP	AR
Poseur	33.26 M	4.52 G	73.2	78.9
连接融合(Concat)	50.94 M	5.06 G	74.7	80.0
相乘融合	50.94 M	5.06 G	75.1	80.2
GLF-Net(相加融合)	50.94 M	5.06 G	**75.3**	**80.4**

综上,对于人体姿态估计主干网络来说,上述结果均证明了全局特征与局部特征进行相加融合对于特征提取上的有效性。GLF-Net 通过综合利用全局极化自注意力和局部深度可分离卷积,能够更好地捕捉不同兴趣区域和关键点之间的结构信息,从而提高了对复杂人体姿态的准确性。全局极化自注意力使模型能够关注整体上的关键点分布,而局部深度可分离卷积则有助于细化关键点的特征表示,两者相互协同,使得 GLF-Net 在姿态估计任务上表现更为卓越。此外,通过观察各模型的算力和参数量,可以发现 GLF-Net 在取得更好性能的同时,在有限的设备条件下,尽管显著地增加了计算负担,该模型在实际应用中仍具有一定的可行性。

2.5　可视化结果

为表明改进后的算法 GLF-Net 在人体姿态估计中因遮挡、视角、人员密集造成的部分部位重叠等影响下,与有一定的抗干扰能力,本文在 COCO 数据集上进行了可视化,并于原网络 Poseur 的可视化结果进行了对比,以此对全局-局部特征融合网络的特征提取增强效果进行说明。如图 8 所示,包含了单人、多人、遮挡、不同视角以及部位重叠的人体姿态估计结果,其中图 8(a)是单人遮挡下的姿态估计;图 8(b)是多人遮挡下的姿态估计;图 8(c)是多人在不同视角下的姿态估计;图 8(d)是多人密集站位导致部分关键点重叠的姿态估计。从图中可以看出,Poseur 和 GLF-Net 模型不论是在单人还是多人的场景下都能进行人体姿态估计,但 GLF-Net 在遮挡或重叠等影响下,能够对检测出的关键点进行正确的建模,并对建模错误的关键点进行修正,具有较好的抗干扰能力。这表明本文所提算法具有一定的优越性,可以在大部分情况下保持良好的人体姿态估计性能。

(a) 单人+遮挡　(b) 多人+遮挡　(c) 多人+视角　(d) 多人+部位重叠

图 8　姿态图像可视化结果 1

然而,在某些特定场景下该算法也具有一定的局限性,如图 9 列举了部分失败的例子。图 9(a)为人体动作快速变化时产生的运动模糊,导致图像分辨率较低,模型可能无法实时跟踪到所有关键点,使得姿态估计错误;图 9(b)是在复杂场景下,由于人体姿态非常多样,模型表现不佳,可能会导致姿态估计的困难;图 9(c)是由于模型对光照、背景和相机距离等因素更为敏感,可能会造成如图中的左右腿关键点识别错误的问题;图 9(d)是由于人体在不同的距离下会呈现不同的尺度,同时也会受服装风格等因素的影响,可能无法捕捉到局部关键点的准确位置。

(a) 运动模糊　(b) 复杂背景

(c) 光照因素　(d) 尺度变化

图 9　姿态图像可视化结果 2

3　结　　论

为了解决骨干网络 CNN 在人体姿态估计任务中因特征提取不足而导致特征信息丢失的问题，本文提出了基于全局-局部特征的人体姿态估计算法 GLF-Net，旨在探究双支路联合局部深度可分离卷积与全局极化自注意力机制进行特征信息高效融合的有效性。该算法通过增添全局-局部特征融合模块，使得整个模型更加关注人体区域的全局和局部特征信息，实现网络结构的深化和特征提取的持续增强，有效解决了骨干网络在特征提取时造成的信息丢失问题，提升了人体关键点检测的准确性，获得较好的姿态估计效果。然而，本文算法也具有一定的不足，当人体受到尺度变化、运动模糊、关照因素和姿势复杂性等因素的影响下，姿态估计的鲁棒性可能会受到挑战。未来在后续的实验中希望通过综合使用多种技术手段，不断改进和优化算法，考虑模型的轻量化问题，使其在更广泛的场景下取得更好的性能。

参考文献

[1] 乔迤，曲毅. 基于卷积神经网络的 2D 人体姿态估计综述[J]. 电子技术应用，2021，47(6)：15-21.

[2] 秦晓飞，郭海洋，陈浩胜，等. 基于深度残差网络的多人姿态估计[J]. 光学仪器，2021，43(2)：39-47.

[3] XIAO B，WU H，WEI Y. Simple baselines for human pose estimation and tracking[C]. Proceeding of the 15th European Conference on Computer Vision(ECCV). September 8-14，2018，Munich，Germany：Springer，2018：472-487.

[4] YU C，XIAO B，GAO C，et al. Lite-hrnet：A lightweight high-resolution network[C]. Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition，Nashville，TN，USA：IEEE Computer Society Press，2021：10435-10445.

[5] 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.

[6] 梁桥康，吴樾. 基于 HRNet 的轻量化人体姿态估计网络[J]. 湖南大学学报(自然科学版)，2023，50(2)：112-121.

[7] GONG X，CHEN W，JIANG Y，et al. AutoPose：Searching multi-scale branch aggregation for pose estimation[J]. Computing Research Repository，ArXiv preprint arXiv：2008.07018，2020.

[8] CHENG B，XIAO B，WANG J，et al. HigherHRNet：Scale-aware representation learning for bottom-up human pose estimation[J]. Proceedings of 2020 IEEE Computer Society Conference on Computer Vision and Pattern Recognition，Seattle，WA，USA：IEEE Computer Society Press，2020：5385-5394.

[9] LI Q，ZHANG Z，XIAO F，et al. Dite-hrnet：Dynamic lightweight high-resolution network for human pose estimation[C]. Proceedings of the International Joint Conference on Artificial Intelligence. November 6-9，2022，Shenzhen，China：Morgan Kaufmann，2022：1095-1101.

[10] 石跃祥，许湘麒. 基于改进 DenseNet 网络的人体姿态估计[J]. 控制与决策，2021，36(5)：1206-1212.

[11] MA H，WANG Z，CHEN Y，et al. PPT：Token-pruned pose transformer for monocular and multi-view human pose estimation[C]. Proceeding of the 17th European Conference on Computer Vision(ECCV). October 23-27，2022，Tel Aviv，Israel：Springer，2022：424-442.

[12] REN F. Distilling token-pruned pose transformer for 2D human pose estimation[J]. Computing Research Repository，ArXiv preprint arXiv：2304.05548，2023.

[13] MNIH V，HEESS N，GRAVES A，et al. Recurrent models of visual attention[J]. Conference on Neural Information Processing Systems，2014，27(1)：2204-2212.

[14] HUA G，LI L，LIU S. Multipath affinage stacked—hourglass networks for human pose estimation[J]. Frontiers of Computer Science，2020，14(4)：219-229.

[15] BAO Y，ZHANG M，GUO X. Human pose estimation based on improved high resolution network[J]. Journal of Physics：Conference Series，2021，1961(1)：12-60.

[16] OU J，CHEN M，WU H. Full-resolution encoder-decoder networks with multi-scale feature fusion for human pose estimation[J]. Computing Research Repository，ArXiv preprint arXiv：2106.00566，2021.

[17] ZHANG C，HE N，SUN Q，et al. Human pose estimation based on attention multi-resolution network[J]. International Conference on Multimedia Retrieval(ICMR)，2021：682-687.

[18] 方芹，缪宁杰，张如宏，等. 基于注意力机制的人体姿态估计网络[J]. 机械设计与制造工程，2022，51(3)：117-122.

[19] YE S，ZHANG Y，HU J，et al. DistilPose：Tokenized pose regression with heatmap distillation[C]. Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition，Vancouver，BC，Canada：IEEE Computer Society Press，2023：2163-2172.

[20] VASWANI A，SHAZEER N，PARMAR N，et al. Attention is all you need[J]. Advances in Neural

Information Processing Systems, 2017, 30(1): 5998-6008.

[21] SUN G, YE C, WANG K. Focus on what's important: self-attention model for human pose estimation[J]. Computing Research Repository, ArXiv preprint arXiv: 1809.08371, 2018.

[22] LI Y, ZHANG S, WANG Z, et al. TokenPose-learning keypoint tokens for human pose estimation[C]. Proceeding of 2021 IEEE/CVF International Conference on Computer Vision(ICCV), Montreal, QC, Canada: IEEE Computer Society Press, 2021: 11293-11302.

[23] YUAN Y, FU R, HUANG L, et al. HRFormer: High-resolution transformer for dense prediction[J]. Computing Research Repository, ArXiv preprint arXiv: 2110.09408, 2021.

[24] XU Y, ZHANG J, ZHANG Q, et al. ViTPose: Simple vision transformer baselines for human pose estimation[J]. Computing Research Repository, ArXiv preprint, arXiv: 2204.12484, 2022.

[25] LI L, ZHAO L, XU L, et al. Towards high performance one-stage human pose estimation[J]. ACM Multimedia Asia, 2023, 37: 1-5.

[26] ZHOU Y, WANG X, XU X, et al. X-hrnet: Towards lightweight human pose estimation with spatially unidimensional self-attention[C]. Proceeding of 2022 IEEE International Conference on Multimedia and Expo(ICME), Taipei, Taiwan: IEEE Computer Society Press, 2022: 1-6.

[27] CHEN X, WANG H, NI B. X-volution: On the unification of convolution and self-attention[J]. Computing Research Repository, ArXiv preprint arXiv: 2106.02253, 2021.

[28] PAN X, GE C, LU R, et al. On the integration of self-attention and convolution[C]. Proceeding of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA: IEEE Computer Society Press, 2022, 2022(1): 805-815.

[29] ZHAO S, LIU K, HUANG Y, et al. DPIT: Dual-pipeline integrated transformer for human pose estimation[J]. CAAI International Conference on Artificial Intelligence. Cham: Springer Nature Switzerland, 2022, 13605: 559-576.

[30] MAO W, GE Y, SHEN C, et al. Poseur: Direct human pose regression with transformers[C]. Proceeding of the 17th European Conference on Computer Vision(ECCV). October 23-27, 2022, Tel Aviv, Israel: Springer, 2022, 13666: 72-88.

[31] LIU H, LIU F, FAN X, et al. Polarized self-attention: towards high-quality pixel-wise regression[J]. Computing Research Repository, ArXiv preprint arXiv: 2107.00782, 2021.

[32] CHOLLET F. Xception: Deep learning with depthwise separable convolutions[C]. Proceeding of 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA: IEEE Computer Society Press, 2017: 1800-1807.

[33] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft coco: Common objects in context[J]. Lecture Notes in Computer Science, 2014, 8693: 740-755.

[34] ANDRILUKA M, PISHCHULIN L, GEHLER P, et al. 2D human pose estimation: new benchmark and state of the art analysis[C]. Proceeding of 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA: IEEE Computer Society Press, 2014: 3686-3693.

作者简介

毛琳,博士,副教授,主要从事多传感器特征融合与目标跟踪方面的研究。

任春贺,硕士研究生,主要从事计算机视觉图像处理与2D人体姿态估计方面的研究。

E-mail: renchunhe7241@163.com

杨大伟,博士,副教授,主要从事图像处理与计算机视觉方面的研究。