基于隐式关键点互联的人体姿态估计矫正器算法

才 华[1],朱瑞昆[1],付 强[2],王伟刚[3],马智勇[3],孙俊喜[4]

(1.长春理工大学 电子信息工程学院,长春 130022;2.长春理工大学 空间光电技术研究所,长春 130022;3.吉林大学第一医院 泌尿外二科,长春 130061;4.东北师范大学 信息科学与技术学院,长春 130117)

摘 要:针对现有人体姿态估计模型中忽略人体关键点之间相互联系的问题,提出了一种基于相关矩阵建立隐式关键点互联的人体姿态估计矫正算法。该算法采用双阶段网络结构,第一阶段使用现有模型获得关键点的准确率,第二阶段计算相关系数矩阵,构建关键点矫正器,实现隐式的关键点互联,通过选取效果较好的部分关键点作为指导,优化效果较差的另一部分关键点。将伪热力图信息重组后分为指引区域和待索引区域两个矩阵,指引矩阵根据准确率分为3个区域,并和转置后的待索引矩阵分别相乘得到相关矩阵,为不同相关矩阵增加不同的影响因子,通过公式得到总相关矩阵,与待索引矩阵相乘得到索引区域,最终逆向信息重组恢复到最初维度,与伪热力图进行逐像素相加,得到最终预测热力图,自动建立各关键点之间的联系,矫正现有模型忽略的关键点相互关系。在COCO2017数据集上的实验结果表明:本文人体姿态估计矫正算法在关键点识别准确度为70.5%,比现有模型SimpleBaseLine平均提升了0.4%。在一些容易被遮挡的部位,例如脚踝和手腕,体积小,处于四肢末端,遮挡的概率高,相邻的关键点个数少,表层信息特征少,人体姿态估计矫正器通过加强四肢末端关键点和其他全身关键点的联系,建立更多的特征联系。实验表明:脚踝和手腕识别的准确率相对提升1.5%,相比于其他现有模型,准确率也有较大的提升,证明了隐式关键点互联的人体姿态估计矫正算法的有效性。

关键词:机器视觉;人体姿态估计;隐式关键点;相关矩阵

中图分类号:TP391.4 **文献标志码**:A **文章编号**:1671-5497(2025)03-1061-11

DOI:10.13229/j.cnki.jdxbgxb.20230600

Human pose estimation corrector algorithm based on implicit key point interconnection

CAI Hua[1],ZHU Rui-kun[1],FU Qiang[2],WANG Wei-gang[3],MA Zhi-yong[3],SUN Jun-xi[4]

(1. *School of Electronic Information Engineer*, *Changchun University of Science and Technology*, *Changchun* 130022, *China*; 2. *School of Opto-Electronic Engineer*, *Changchun University of Science and Technology*, *Changchun* 130022, *China*; 3. *No.*2 *Department of Urology*, *The First Hospital of Jilin University*, *Changchun* 130061, *China*; 4. *College of Information Science and Technology*, *North Normal University*, *Changchun* 130117, *China*)

收稿日期:2023-06-13.

基金项目:国家自然科学基金重大项目(61890963);吉林省科技发展计划项目(20210204099YY).

作者简介:才华(1977-),男,副教授,博士.研究方向:机器视觉.E-mail:caihua@cust.edu.cn

Abstract: To solve the problem of ignoring the interrelation between key points in the existing human pose estimation models, a new algorithm based on correlation matrix to establish implicit key point interconnection was proposed. The algorithm adopts a two-stage network structure. In the first stage, the existing model was used to obtain the accuracy of key points. In the second stage, the correlation coefficient matrix was calculated to build a key point corrector to achieve implicit key point interconnection. By selecting some key points with good effect as guidance, the other key points with poor effect were optimized. The pseudo-thermal map information was reorganized into two matrices, the guide region and the region to be indexed. The guide matrix was divided into three regions according to the accuracy rate, and was multiplied by the transposed matrix to be indexed to obtain the correlation matrix, and different influence factors were added to different correlation matrices. The total correlation matrix was obtained by formula, and the index region was obtained by multiplying the matrix to be indexed. Finally, the reverse information was reconstructed to the original dimension, and the pseudo-thermal map was added pixel by pixel to obtain the final predicted thermal map, which automatically establishes the relationship between key points and corrects the relationship between key points ignored by the existing model. The experimental results on COCO2017 data set show that the key point recognition accuracy of the human pose estimation correction algorithm is 70.5%, an average increase of 0.4% compared with the existing model SimpleBaseLine. In some parts that are easily obscured, such as ankles and wrists, they are small in size and at the end of limbs, and the probability of occlusion is high. The number of adjacent key points is less, and the surface information features are less. The human posture estimation corrector strengthens the connection between the key points at the extremities and other key points in the whole body, and establishes more feature connections. The experiments demonstrate that the recognition accuracy of ankles and wrists has improved by 1.5%, with a significant increase compared to other existing models, proving the effectiveness of the implicit keypoint-interconnected human pose estimation correction algorithm.

Key words: computer vision; human pose estimation; implicit key points; correlation matrix

0 引 言

基于图像的二维人体姿态估计处理在计算机视觉中占据着重要的地位[1]。人体姿态估计的关键点判断结果会很大程度地影响后续图像中人体动作判断的更深层次分析,例如,无限制场景下的姿态跟踪[2]、人体行为识别[3-5]及预测[6,7]、人机交互[8-10]。因此,提高人体姿态估计网络的性能成为解决实际视觉问题的重要途径之一。根据判断人数的不同,人体姿态估计可以分为单人人体姿态估计[11-13]和多人人体姿态估计[14-16],单人人体姿态估计只能识别仅有一人的图像或者视频,使用较少,多人人体姿态估计由于对可识别人数不受限制,在日常的实际应用更为广泛。

根据识别过程的不同,人体姿态估计又可以分成自顶向下和自底向上的人体姿态估计算法。自顶向下的人体姿态估计算法[17-19]是一种基于目标检测的方法,先使用目标检测技术检测出图像中的人体,然后将每个人体分割成若干个单人目标图像,再对每个单人图像进行关键点检测以得到最终的姿态估计结果。自顶向下的多人姿态估计的处理方法主要分为两个步骤:首先,采用多人目标检测方法检测到图像中所有人的所在位置,并形成检测框;其次,将多人目标分为多个单人目标图像,进行单人人体姿态估计处理,进行最终的关键点检测。这种方法可以有效地避免多人姿态判断时存在相互遮挡和重叠等问题的影响,从而提高姿态估计的准确率。自底向上的姿态估计算法[20-22]是从图像中每个像素出发,逐层构建出一些可以组成人体姿态的关键点,然后利用这些关键点构建出完整的人体姿态。这种方法不需要事先知道人体的位置和数量,能够对多个人体同时进行姿态估计。

在多人姿态估计领域,由于存在遮挡和重叠等问题,自底向上的姿态估计算法[23-26]通常面临着较大的困难并且准确率较低,很难准确地检测到所有人体姿态。相反,自顶向下的算法采用目

标检测技术可以有效避免遮挡问题,提高识别准确率。然而,对于多个人体同时出现在同一张图像中的情况,自顶向下算法也可能存在漏检和误检的问题。因此,提高多人姿态估计效果还需要进一步研究。目前,许多研究都集中于通过高精度的关键点检测算法提高姿态估计的准确率,使用深度学习模型和注意力机制等技术,通过采用多层次的特征提取和优化算法等方法,有效提高人体姿态估计的效率和准确率[18]。

在复杂环境下的人体姿态估计,会受到很多因素的影响,包括复杂不确定的环境、颜色多变的人体服装、人体关键点被遮挡和图像拍摄角度问题,导致人体关键点不可见和预测难度大,极大地挑战算法处理判断。单人姿态估计使用的范围较小,不适用于多数复杂场景,自底向上的检测方式相比于自顶向下的检测方式在多人姿态估计中效果较差,因此,本文采用基于关键点互联的人体姿态估计新方法,通过在传统人体姿态估计网络的基础上增加关键点互联矫正器,建立相关系数矩阵,挖掘关键点之间的深层次联系,使用较好效果的部分关键点作为指导,优化效果较差的另一部分关键点,自动建立各关键点之间的联系,矫正现有模型忽略的关键点相互关系,自顶向下检测方式可实现多人姿态估计,结合关键点互联的网络矫正,解决传统网络中忽视关键点之间的相互联系,没有针对性地挖掘关键点更深层次的关系,可以克服复杂环境中不稳定的环境和人体的相互遮挡,挖掘更多的特征信息。

1　相关工作

人体姿态估计是计算机视觉领域的重要研究方向,它可以应用于运动分析、人机交互、虚拟现实、医学等领域。近年来,基于深度学习的人体姿态估计方法取得了巨大的进展,其中基于热图的方法备受关注。

在基于热图的人体姿态估计算法中,常见的思路是将关键点与目标图像上的像素一一对应,通过学习关键点与像素之间的对应关系,估计人体骨骼姿态。Hourglass网络[27]是一种常用于人体姿态估计的神经网络结构,其特点是堆叠多个Hourglass模块,预测热图和协方差矩阵,将多尺度的信息结合起来进行姿态估计。具有multi-task模块的CPN[28]包含全局姿态估计和细节姿态再次估计两个部分,通过设计两个堆叠的网络,分别用于明显关键点的检测和部分困难关键点的检测,两个网络分工明确,通过分层次处理完成人体关键点检测。该网络采用了残差连接和注意力机制,能够在处理复杂环境和多人场景时获得较高的准确度。HRNet[29]是一种高分辨率的神经网络,采用多分辨率信息进行特征提取,不断进行多尺度特征信息统合,采用并行的高分辨率和低分辨率子网进行人体姿态预测,能够有效提高估计精度和鲁棒性。Simple Baseline[30]是一种基于ResNet网络结构的姿态估计模型,其特点是引入多尺度训练和测试技术,同时将姿态估计问题转换为每个关键点的检测问题,从而提高准确度和鲁棒性,主干网络后连接逆卷积模块进行热力图预测,获得分辨率特征图像和热力图,最终根据热力图得到人体关键点结果。这些方法只注意从全局的角度上进行提取特征,并没有考虑到人体关键点之间独特的相互关系,遗漏了人体总体的协调信息。

基于关键点互联的姿态估计是近年来受到广泛关注的另一种人体姿态估计方法,它与基于热图的方法不同,基于关键点互联的方法是通过学习关键点之间的联系,估计人体骨架姿态,能够有效解决人体间相互遮挡、变形等问题。Pose Machine[31]是一种基于关键点的神经网络模型,结合了卷积神经网络(CNN)[32]和人造神经网络(FNN)[33]。该网络采用多阶段递归推理流程,并建立了一个关键点互联图估计人体骨架姿态,通过不断迭代更新预测结果,提高准确性和鲁棒性。OpenPose[34]是另一种常用的基于关键点互联的人体姿态估计算法,通过在两个相邻关键点之间建立一个有向场进行匹配,匹配成功就认为是同一个人的关节。以此类推,对所有相邻点做此匹配操作,最后就得到每个人的所有关键点。Deep high-resolution[29]也可以被放在这个区域,因为HRNetv2结合了像基于关键点互联的方法和像Hourglass神经网络一样的多尺度特征和分层结构,它在多任务环境下表现优异,推广领域为人体姿态估计、实例分割、目标检测和语义分割。卷积姿态机(Convolutional pose machine,CPM)[31]是一种基于卷积神经网络的人体姿态估计算法,该算法通过结构化边界特征和多阶段回归学习每个关键点的位置,并建立一个关键点图谱有效地处

理变形和遮挡问题。关联性嵌入(Associative embedding,AE)[23]是另一种基于关键点互联的算法,与其他算法不同,它不是预测每个关键点的位置,而是学习关键点之间的关系,通过建立一个二元关键点关联图估计人体骨架姿态。该方法能够高效地解决遮挡问题和多人姿态估计问题。基于关键点互联的人体姿态估计方法相比于基于热图的方法,在解决多人姿态估计、遮挡等问题方面具有优势,能够更好地适应各种不同场景。基于关键点互联的姿态估计方法只考虑了关键点之间的联系对估计算法的影响,过度关注关键点的相互联系,忽略了提取特征形成热力图的主导地位。

基于此,本文结合基于热图和基于关键点的优点,提出一种人体姿态估计矫正算法,通过准确程度较高的关键点指引准确程度较低的关键点进行位置矫正,改进了基线算法,增加了基于关键点互联矫正器,网络结构如图1所示。该网络首先通过现有传统人体姿态网络进行初步的人体姿态估计,得到初步整体热力图。在网络模型后面增加一个基于隐式关键点互联的姿态矫正器,矫正器首先将初步整体热力图进行信息重组,得到包含热力图的所有信息,其次通过对重组信息的二维矩阵进行拆分和计算,将准确率高的关键点区域作为索引区域,将准确率低的关键点区域作为待索引区域,进行索引特征激活,然后进行逆向信息重组,得到特征激活热力图,最终将初步整体热力图和特征激活热力图进行拼接相加操作,得到真实整体热力图。实验结果表明:该隐式关键点互联的矫正器通过相关矩阵提取关键点之间的相互关系,实现提高人体姿态估计中关键点的准确率。

2 隐式关键点互联姿态估计矫正算法

2.1 网络结构

隐式关键点互联姿态估计矫正算法网络结构如图1所示。

利用传统人体姿态估计网络对数据集中的图像进行人体姿态估计,获取人体姿态估计热力图并进行重新排序,排序规则为:将每个识别关键点的估计准确率由高到低排序,形成伪热力图,记为$P=H\times W\times K$,H为伪热力图的高,W为伪热力图的宽,K为热力图的通道,也是网络关键点个数,在COCO2017数据集中,关键点个数为17个,因此,$K=17$;级联金字塔CPN[28]网络提出了一种网络结构将网络分为GlobalNet和RefineNet两个阶段,网络第1阶段GlobalNet只定位容易发现的关键点,无法准确识别遮挡或不可见的关键点,第2阶段RefineNet通过整合来自GlobalNet的特征表示,进行“硬”关键点挖掘。受到CPN的影

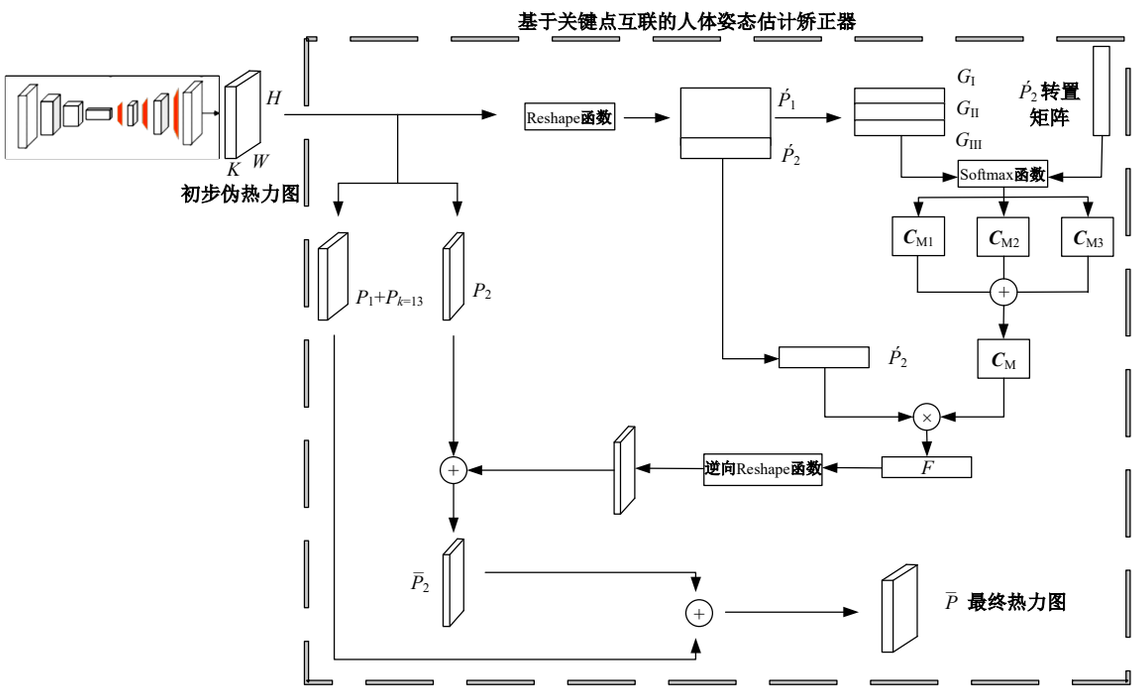

图1 本文网络的详细脉络结构图

Fig. 1 Detailed diagram of the network structure in this paper

响，本文尝试通过建立隐式联系类比于RefineNet所得到的关键点特征信息建立“硬”关键点挖掘，将伪热力图P分为两部分，表示为$P=\{P_1,P_{k=13},P_2\}$，其中，第1部分的伪热力图$P_1=H\times W\times K_1$，$K_1=12$，第2部分的伪热力图$P_2=H\times W\times K_2$，$K_2=4$，进行信息重组；因为后续特征信息激活涉及矩阵的拆分计算，奇数关键点个数会打乱相关矩阵计算，因此，在信息重组过程中，不对$P_{k=13}$进行计算。

信息重组是线性变换映射，将图像转换成为一个二维数组，记为：

$$\dot{P}=\text{Reshape}(H\times W\times K) \tag{1}$$

将二维图像所在区域分为：

$$\dot{P}_1=\text{Reshape}(H\times W\times K_1) \tag{2}$$

和

$$\dot{P}_2=\text{Reshape}(H\times W\times K_2) \tag{3}$$

两个区域，其中，$\dot{P}_1$为指引关键点区域，$\dot{P}_2$为待索引关键点区域，并且$K_1=12$，$K_2=4$；K_1、K_2的选取以及后续的区域划分问题通过实验获得。

为了便于进行矩阵乘法计算，将$\dot{P}_1$从上到下均匀分为3个区域，记为$G_{\text{Ⅰ}}$、$G_{\text{Ⅱ}}$、$G_{\text{Ⅲ}}$，这3个区域属于未知区域，所代表的相关性不确定，其中：

$$G_{\text{Ⅰ}}=HW\times K_1^{\text{Ⅰ}} \tag{4}$$

$$G_{\text{Ⅱ}}=HW\times K_1^{\text{Ⅱ}} \tag{5}$$

$$G_{\text{Ⅲ}}=HW\times K_1^{\text{Ⅲ}} \tag{6}$$

式中：$K_1^{\text{Ⅰ}}=\{1\leqslant K\leqslant 4\}$，$K_1^{\text{Ⅱ}}=\{5\leqslant K\leqslant 8\}$，$K_1^{\text{Ⅲ}}=\{9\leqslant K\leqslant 12\}$。

利用分区后的$\dot{P}_1$计算并获得指引关键点区域$\dot{P}_1$与待索引关键点区域$\dot{P}_2$之间的3个相关系数矩阵，相关系数矩阵代表指引关键点区域和待索引关键点区域之间的相关性，3个相关系数矩阵分别为

$$\boldsymbol{C}_{\text{M1}}=\text{Softmax}(G_{\text{Ⅰ}}\bullet\dot{P}_1^{\text{T}}) \tag{7}$$

$$\boldsymbol{C}_{\text{M2}}=\text{Softmax}(G_{\text{Ⅱ}}\bullet\dot{P}_2^{\text{T}}) \tag{8}$$

$$\boldsymbol{C}_{\text{M3}}=\text{Softmax}(G_{\text{Ⅲ}}\bullet\dot{P}_2^{\text{T}}) \tag{9}$$

因为3个区域的准确率依次下降，3个相关矩阵系数对总相关系数矩阵的影响也不同，所以给3个相关矩阵系数赋予不同的比例系数再相加，进而获得总相关系数矩阵$\boldsymbol{C}_{\text{M}}$：将分布特征表示映射到样本标记空间，为每个区域找到适合区域本身的影响因子，设计了一个全连接层，输入层为：

$$P=\text{Reshape}(H\times W\times K_1) \tag{10}$$

输出层只有3个元素e^{M1}、e^{M2}、e^{M3}，分别对应$\boldsymbol{C}_{\text{M1}}$、$\boldsymbol{C}_{\text{M2}}$、$\boldsymbol{C}_{\text{M3}}$的影响因子。

$$\boldsymbol{C}_{\text{M}}=\frac{\boldsymbol{C}_{\text{M1}}}{e^{\text{M1}}}+\frac{\boldsymbol{C}_{\text{M2}}}{e^{\text{M2}}}+\frac{\boldsymbol{C}_{\text{M3}}}{e^{\text{M3}}} \tag{11}$$

使用相关系统矩阵$\boldsymbol{C}_{\text{M}}$对待索引关键点区域$\dot{P}_2$进行矩阵相乘计算，得到索引区域：

$$F=\boldsymbol{C}_{\text{M}}\bullet\dot{P}_2 \tag{12}$$

将索引区域F通过线性变换映射进行逆向信息重组得到$\bar{F}$，此处的信息重组是信息重组结构的逆过程，$\bar{F}$表示由相关系数矩阵$\boldsymbol{C}_{\text{M}}$激活的姿态特征图像，$\bar{F}=H\times W\times K_2$，$K_2=4$。

将姿态特征图像$\bar{F}$与初步伪热力图的第2部分P_2进行逐像素相加操作，获得已索引关键点区域$\bar{P}_2$：

$$\bar{P}_2=\bar{F}\oplus P_2 \tag{13}$$

将第一部分的伪热力图P_1和已索引关键点区域$\bar{P}_2$进行合并，最终得到最终预测热力图$\bar{P}$，$\bar{P}=\{P_1,P_{k=13},\bar{P}_2\}$。

利用训练集对基于关键点互联的被遮挡人体姿态估计矫正器(Key point correction，KPC)以及传统人体姿态估计网络进行端到端训练，在达到指定训练次数后，取损失函数值最小的损失函数作为矫正器在该传统人体姿态估计网络的损失函数，基于关键点互联的被遮挡人体姿态估计矫正器以及传统人体姿态估计网络训练完成。

2.2　注意力信息重组

信息重组是为了提取到人体姿态估计热图中的高级别表示特征，将不同的特征进行重新组合。

本文中的信息重组是将传统人体姿态估计网络中所输出的三维热力图转换成包含特征信息的二维数组，该二维数组包含三维热力图的所有信息(见图2)。三维热力图一共有K个通道，代表每个关键点的热图信息，将每个通道进行全连接展开成一个一维向量，即K个一维向量，在这里设计一个初步判断准确率损失函数，然后根据每个关键点的准确率由高到低进行排列，实现了三维热图转换成为二维数组，再进行高斯模糊处理。

传统人体姿态网络模型输出的初步预测热力图$P=H\times W\times K$，H为伪热力图的高，W为热力图的宽，K为热力图的通道数，也是人体关键点个数，在COCO 2017数据集上，$K=17$；通过全连接的线性展开$\bar{P}=HW\times K$，即K个一维向量，根据初步预测损失函数Loss_1，将K个一维向量按照精准度进行从高到底排序，然后拼接成为二维图

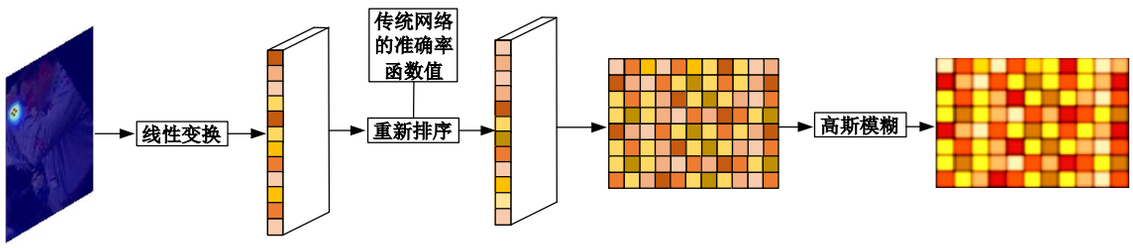

图2 信息重组设计流程

Fig. 2 Information repackaging design process

像。认为此时的热力图变换图像已经包含初步热力图的所有信息,包括结构和特征等信息,为了进一步突出主体信息,采用高斯模糊的方式,对重构后的图像进行处理。计算如下:

$$G(u,v)=\frac{1}{2\pi\sigma^2}\mathrm{e}^{-(u^2+v^2)/(2\sigma^2)} \tag{14}$$

逆向信息重组是将已完成特征激活的索引关键点区域二维数组恢复成深度图像,逆向信息重组的过程结构与信息重组的过程结构相反,利用全连接层将包含特征信息的二维数组$F=HW\times K_2$逆向转变为三维热力图$\bar{F}=H\times W\times K_2$,完成三维热力图的信息融合。

2.3 损失函数

本文设置3个损失函数:第一个损失函数作为注意力信息重组的初步判断,对整体的算法网络影响较小;第二个损失函数是最终预测热力图与真实标签热力图进行比较的最终判断。这两个损失函数都实时参与训练,第二个损失函数作为主要判断模型训练过程的损失函数。利用预测热力图以及标签热力图建立损失函数loss_1和loss_2,loss_1作为初步热力图重构排序的判断依据和训练结果的有限判断依据,通过一定的权重参数影响网络训练,loss_2作为训练结果的主要判断依据,获得相应的损失函数值,Loss是总损失函数,通过联合loss_1和loss_2,进行总体判断。

$$\mathrm{loss}_1=\frac{1}{k}\sum_1^k l(P-\hat{P})^2 \tag{15}$$

式中:P为得到的初步预测值;$\hat{P}$为图像的实际标注值;l为评价系数;K为关键点个数。

$$\mathrm{loss}_2=\frac{1}{k}\sum_1^k l(\bar{P}-\hat{P})^2 \tag{16}$$

式中:$\bar{P}$为得到的最终预测值;$\hat{P}$为图像的实际标注值;l为评价系数;K为关键点个数。

$$\mathrm{Loss}=\mathrm{loss}_2+\lambda\mathrm{loss}_1 \tag{17}$$

式中:Loss为最终的损失函数值;λ为权重系数,经过测试,$\lambda=0.2$时,效果最佳。

3 实验分析

本节中,在COCO2017数据集上进行了大量实验,并与现阶段最先进的方法进行比较,以验证矫正器的性能,所有实验都在GPU为Nvidia GeForce GTX 2080Ti的实验设备上进行,使用Pytorch框架实现。

3.1 实验数据集及评价指标

在COCO2017数据集[34]中,关键点检测任务是标注人体姿态估计的一部分。该任务的目标是检测图像中人体的关键点,例如肩膀、膝盖、手腕、鼻子等部位的位置。COCO2017关键点数据集为1 000多个人类对象,提供了超过20万个关键点标注。这些标注用于描述图像中人物的姿态、行为和动作,对于人体行为分析和动作识别具有重要意义。COCO2017关键点数据集的标注包括17个关键点位置。这些点位置用坐标形式标注在图像中,使用了一种称为“密集人体关键点”的标注方法。

OKS是目前常用的人体姿态关键点检测算法的评价指标,作为本文的一种评价指标,目的是计算预测值和实际值的相似度。

$$\mathrm{OKS}_p=\frac{\sum_i \exp\{-d_{pi}^2/2S_p^2\sigma_i^2\}\delta(V_{Pi}>0)}{\sum_i \delta(V_{Pi}>0)} \tag{18}$$

式中:p为当前图片所有真实行人中ID为p的人,$p\in(0,M)$;M为当前图中共有行人的数量;i为ID为i的关键点;d_{pi}表示当前检测的一组关键点中ID为i的关键点与groundtruth行人中ID为p的人的关键点中ID为i的关键点的欧氏距离;$d_{pi}=\sqrt{(\dot{x}_i-x_{pi})(\dot{y}_i-y_{pi})}$,$(\dot{x}_i,\dot{y}_i)$为当前的关键点检测结果,$(x_{pi},y_{pi})$为真实值;$S_p$为groundtruth行人中ID为$p$的人的尺度因子,其值为行人检测框面积的平方根:$S_p=\sqrt{wh}$,$w$、$h$为检测框的宽和高;$\sigma$为ID为$i$类型的关键点归一化因子,这个因子是

通过对所有的样本集中的关键点由人工标注与真实值存在的标准差,σ越大表示此类型的关键点越难标注。

对COCO 2017数据集中的5 000个样本统计出17类关键点的归一化因子,σ的取值为:{鼻子:0.026,眼睛:0.025,耳朵:0.035,肩膀:0.079,手肘:0.072,手腕:0.062,臀部:0.107,膝盖:0.087,脚踝:0.089},因此,此值可以视为常数。

AP平均准确率也是较为常用的一种评价指标,因为采用自顶向下的检测方式,每次检测图片只会标注一个人体,AP的计算方式为:

$$\mathrm{AP}=\frac{\sum_{P}\delta(\mathrm{OKS}_p>T)}{\sum_{P}1} \tag{19}$$

3.2　消融实验

消融实验是为了得到上述算法中指引区域矩阵$\boldsymbol{K}_1$、待索引区域中$\boldsymbol{K}_2$的值以及指引区域矩阵块划分的值。通过实验发现指引区域矩阵值必须大于待索引区域矩阵值,即$\boldsymbol{K}_1>\boldsymbol{K}_2$,否则人体姿态估计矫正器不能矫正现有模型,甚至起到相反作用,$\boldsymbol{K}_1<\boldsymbol{K}_2$情况,后续的矩阵相乘计算需要舍弃大量的数据。表1是不同的$\boldsymbol{K}_1$、$\boldsymbol{K}_2$以及指引区域矩阵块值对应的实验结果,该结果均在Simple-BaseLine现有模型基础上加上不同值对应的人体姿态估计矫正器上进行。模型统一用SBL-KPC表示,指引区域矩阵块用$\boldsymbol{Q}$表示。

后续的算法中存在矩阵乘法计算,$\boldsymbol{K}_1$只能是$\boldsymbol{K}_2$的整数倍,多余的值只能舍弃,而区域块值$\boldsymbol{Q}$是根据不同的$\boldsymbol{K}_1$、$\boldsymbol{K}_2$值确定的。从表1的检测结果实验数据可以得到随着$\boldsymbol{K}_1$、$\boldsymbol{K}_2$值的变化,模型的准确率呈类抛物线变化,$\boldsymbol{K}_1=13$、$\boldsymbol{K}_2=4$时,准确效果达到最佳。

表1　不同$\boldsymbol{K}_1$、$\boldsymbol{K}_2$、$\boldsymbol{Q}$对应的检测效果

Table 1　Detection results corresponding to different $\boldsymbol{K}_1$,$\boldsymbol{K}_2$ and Q

模型	$\boldsymbol{K}_1$	$\boldsymbol{K}_2$	$\boldsymbol{Q}$	AP/%
SBL-KPC	9	8	1	70.2
SBL-KPC	10	7	1	70.1
SBL-KPC	11	6	1	68.9
SBL-KPC	12	5	2	70.3
SBL-KPC	13	4	3	70.5
SBL-KPC	14	3	4	70.4
SBL-KPC	15	2	7	70.3
SBL-KPC	16	1	16	70.1

3.3　对比实验

对比实验是为了验证所设计矫正器的有效性,评估矫正器如何影响整体模型的性能。因此,进行了大量的对比测试,验证矫正器在不同模型上的性能。在COCO 2017数据集上,本文评估了验证集中所有的图像。将COCO 2017数据集中表现良好的模型方法都列举出来,并在模型的后端加上设计的关键点矫正器,进行多次训练和验证,将不同方法在COCO 2017数据集上的验证确定性对比结构记录在表2中,表2的结果表明本文设计的关键点矫正器可以提升现有模型的关键点准确率判断,代价只是增大一点计算量。从人体所有关键点整体角度出发,矫正器的增加,提升了整体的关键点精度。

表2　不同方法在COCO 20177数据集的姿态估计评价结果

Table 2　Different methods of attitude estimation in COCO 2017 data sets evaluate the results

模型	GFLOPs	AP/%	AP^{50}/%	AP^{75}/%
CPN	9.142	68.4	87.5	74.4
CPN-KPC	9.321	68.7(↑0.3)	87.9(↑0.4)	74.6(↑0.2)
SBL	7.671	70.1	91.3	76.9
SBL-KPC	7.862	70.5(↑0.4)	91.5(↑0.2)	77.1(↑0.2)
HRNeT	1.533	73.5	92.2	80.3
HRNeT-KPC	1.721	73.6(↑0.1)	92.2	80.6(↑0.3)
UDP	9.432	71.2	91.3	78.2
UDP-KPC	9.622	71.4(↑0.2)	91.2(↓0.1)	78.3(↑0.1)
DarkPose	9.452	71.3	90.9	78.4
DarkPose-KPC	9.638	71.4(↑0.1)	91.2(↑0.3)	78.5(↑0.1)

矫正器模块有效性消融研究验证了所设计的矫正器模块有效性。测试未配置矫正器模块的现有模型,记为CPN,其次是在现有模型上增加矫正器模块,记为CPN-KPC,相比于CPN,CPN-KPC的AP、AP^{50}、AP^{75}提升了0.3%、0.4%、0.2%,证明了设计的基于隐式关键点互联的关键点矫正器的有效性。此外,为了验证矫正器在现有模型的普遍适用,对其他主干网络模型也进行了对比测试,分别在SBL、HRNeT、UDP以及DarkPose上增加了关键点矫正器KPC,现有模型的准确率分别提升了0.4%、0.1%、0.2%、0.1%,证明了矫正器的普遍适用性和有效性。本文设计的基于关键点隐式互联矫正器算法是通过准确率较高的关键点指引准确率较低的关键点,进行关键点矫正,可以克服部分关键点遮挡困难,对一些易遮挡关

键点部位的识别准确率,有明显的提升。下面主要是为了验证所设计的方法相比于现有模型的有效性,评估方法在不同的关键点位置的适用性和差异性,因此,进行大量的对比实验。同样地,在COCO 2017数据集中,使用效果较好的现有模型和设计的方法,测试不同方法在不同关键点位置上的表现。将测试结果都列举在表3中,通过表格的方式进行对比,标记了各个关键点中效果最好的数据。

对比实验:首先将数据集中所有的关键点大致分为表3中的几个类别,列举了COCO 2017数据集中几个表现较好的现有模型方法,包括CPN、SBL、HourGlass以及HRNET,在测试集上进行多次重复测验,将得到的结果记录在表3中。使用相同的测试方法对本文设计的方法进行测试,得到的结果也记录在表3中。实验结果对比表明:提升了现有模型网络的性能,在各个关键点位置的估计都有较大的提升,同时这样的改进相比于其他模型,效果提升也很大,尤其是一些容易被遮挡和混淆的关键点,相比于效果较好的HRNeT,改进的方法在脚踝的位置提升了3.6%,相比于SBL,脚踝和手腕准确率分别提升了2.8%和0.9%,对于手腕脚腕等人体四肢末端提升约1.5%,证明本文方法效果较好。图3展示了部分整体热力图结果。

表3 不同方法在各个关键点的估计结果

Table 3 Estimation results of different methods at various key points %

模型	Head(头)	Sho.(肩)	Elb.(肘)	Wri.(腕)	Hip(臀)	Knee(膝)	Ank(踝)	OKS
CPN	90.1	86.2	85.1	83.5	86.1	84.3	80.2	75.2
CPN-KPC	90.2	86.4	85.3	85.2	87.1	85.2	83.1	78.4
SBL	92.2	91.1	89.2	87.8	90.1	87.2	86.3	90.3
SBL-KPC	92.4	91.5	89.5	88.1	90.5	88.3	89.1	91.4
HourGlass	91.8	90.3	88.7	87.3	90.3	88.2	85.3	91.3
HourGlass-KPC	91.8	90.5	89.1	87.6	91.2	89.3	88.2	92.5
HRNeT	92.4	92.1	90.3	89.2	90.4	89.3	86.2	91.8
HRNeT-KPC	92.4	92.2	90.5	91.1	90.8	89.9	89.8	93.2

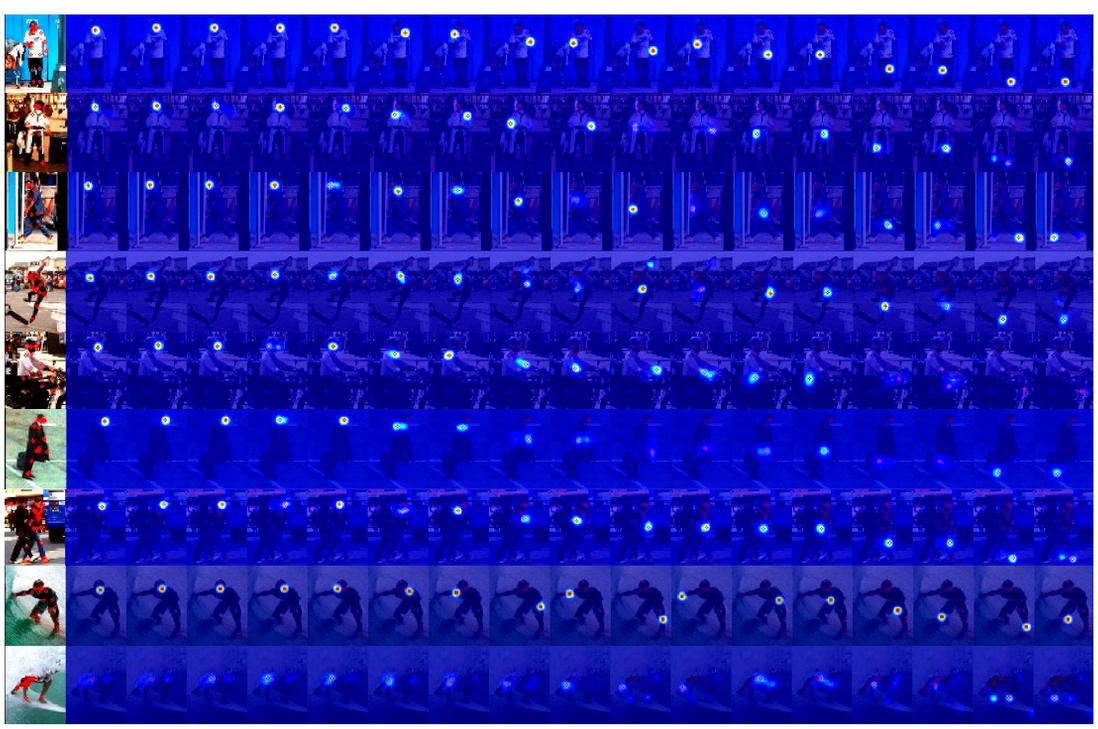

图3 最终预测热力图结果

Fig. 3 Final predicted heat map results

人体姿态热力图是一种以颜色编码的方式展示人体姿态的热成像图,可以直观地反映出人体不同部位的姿态信息。它通过分析关键点在图像中的位置和姿态信息,将不同部位的姿态信息用不同的颜色表示,从而形成一张色彩相互呼应、清晰明了的图像。热力图每行分别是原图像和对应的17个关键点热力图,图像中亮点为预测关键点区域,预测效果较好,通过人体姿态热力图判断关键点预测效果。

为了显示提出的基于隐式关键点互联的人体

姿态估计算法在复杂环境下的鲁棒性,随机挑选了包含人体的测试图片进行关键点预测。图 4 为 COCO 2017 数据集中的人体姿态可视化结果。本文方法的人体姿态估计验证结果如图 4 所示,包括单人图像和多人图像。为了表达清晰,每张图像只标注了多人中一人的关键点估计位置,并且对于关键点个数不足的人体,只估计可视的部分关键点。

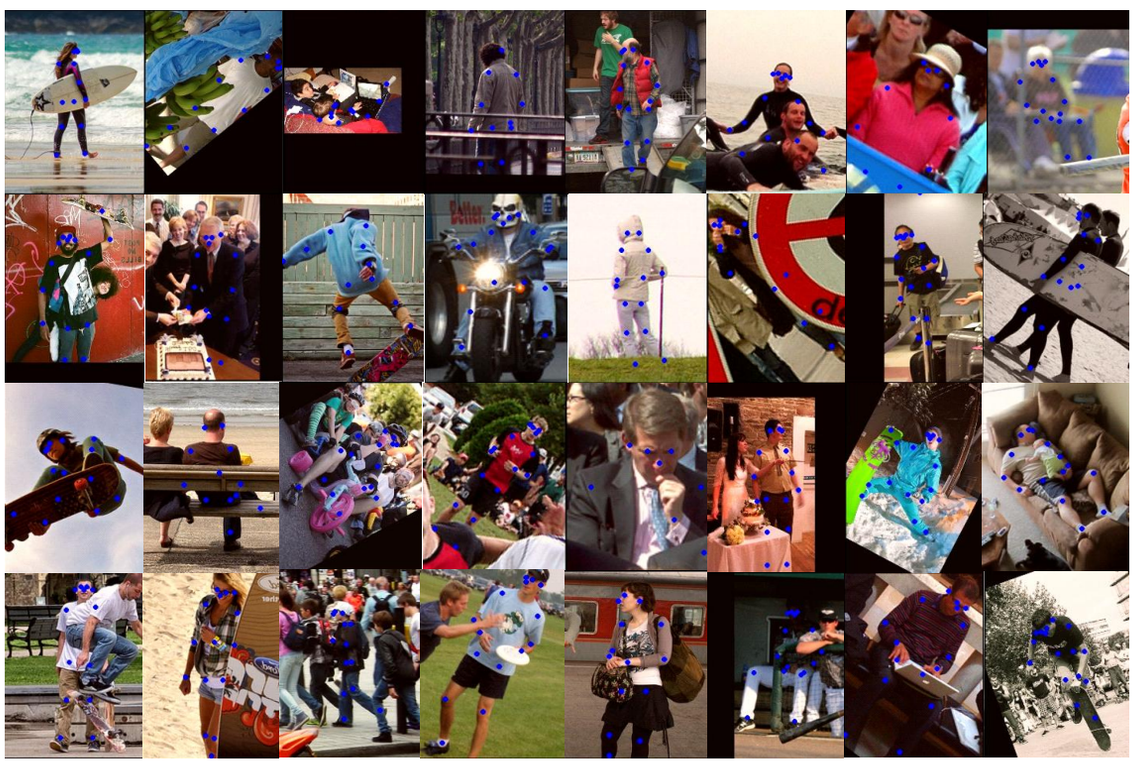

图 4　测试集中部分关键点标注结果

Fig. 4　Annotation results of some key points in the test set

为了证明本文人体姿态估计矫正器算法的有效性,在简单基线(Simple base line,SBL)网络加上矫正器算法进行测试,如图 4 所示,选取了一部分测试结果,图片包括物体遮挡、人体相互遮挡以及复杂环境的情况下进行人体姿态估计。人体关键点部分遮挡情况下,可以完成绝大部分的人体关键点检测。从图 4 中可以看出,本文方法在不同的人物个体中都具有良好的效果,克服了复杂环境和遮挡问题的影响,验证结果表明本方法在人体姿态估计中关键点识别度高。

为了直观验证人体姿态估计矫正器的有效性,本文选取了一部分传统网络检测效果和增加矫正器以后的检测效果。图 5 第 1 行图片展示了未增加矫正器的关键点检测效果,部分关键点受到复杂环境和遮挡的影响,检测效果有偏差;第 2 行则展示了增加矫正器后的关键点检测效果,相

图 5　传统网络检测效果和增加矫正器检测效果

Fig. 5　Effect of traditional network detection and the effect of additional orthotics detection

比于第一行更准确,且与实际位置偏差较小,直观地证明了矫正器有效。

4 结束语

本文提出了一种基于相关矩阵建立隐式关键点互联的人体姿态估计矫正算法,以解决现有模型中忽略人体关键点之间相互联系的问题。实验结果表明,该算法能够自动建立各关键点之间的联系,优化效果较差的关键点,相比现有模型SBL平均提升0.4%,在一些容易被遮挡的部位,例如脚踝和手腕,识别准确率相对提升1.5%,在其他现有模型使用矫正器,检测准确率也有一定的提升。证明提出的人体姿态估计矫正算法是有效的,具有较好的检测效果。

参考文献:

[1] Andriluka M, Iqbal U, Insafutdinov E, et al. PoseTrack: a benchmark for human pose estimation and tracking[C]//IEEE / CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 5167-5176.

[2] Liu Q, Zhang Y, Bai S, et al. Explicit occlusion reasoning for multi-person 3D human pose estimation[C]//Proceedings of the European Conference on Computer Vision (ECCV), Switzerland, 2022: 497-517.

[3] Wu H, Ma X, Li Y. Multi-level channel attention excitation network for human action recognition in videos[J]. Signal Processing: Image Communication, 2023, 114: No. 116940.

[4] Nguyen P, Liu T, Prasad G, et al. Weakly supervised action localization by sparse temporal pooling network[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 2018: 6752-6761.

[5] Ren Z, Zhang Q, Gao X, et al. Multi-modality learning for human action recognition[J]. Multimedia Tools and Applications, 2021, 80(11): 16185-16203.

[6] Tang Y, Liu R. Skeleton embedding of multiple granularity attention network for human action recognition[C]//International Conference on Articulated Motion and Deformable Objects. Berlin: Springer, 2020: 12878-12885.

[7] Liang Z J, Wang X L, Huang R, et al. An expressive deep model for human action parsing from a single image[C]//EEE International Conference on Multimedia and Expo (ICME), Chengdu, China, 2014: 1-6.

[8] Goh E S, Sunar M S, & Ismail A W. 3D object manipulation techniques in handheld mobile augmented reality interface: a review[J]. IEEE Access, 2019, 7: 40581-40601.

[9] Yang X, Chen Y, Liu J, et al. Rapid prototyping of tangible augmented reality interfaces: towards exploratory learning for science education[J]. Interactive Learning Environments, 2019, 27(4): 469-483.

[10] Zhang D, Peng Y, Yang W, et al. Multi-viewpoint interaction with social robots: a case study of speech therapy for children with autism[J]. Journal of Intelligent & Robotic Systems, 2018, 92(3/4): 359-3728.

[11] Zhang R, Li J, Xiao T, et al. BodyPoseNet: body pose estimation driven by deep neural networks[J]. Signal Processing: Image Communication, 2021, 99: No. 116290.

[12] Chen L, Papandreou G, Kokkinos I, et al. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4):834-848.

[13] Zeng W , Gao Y, Zheng Y, et al. DenseReg: fully convolutional dense regression for accurate 3D human pose estimation[J]. IEEE Transactions on Image Processing, 2021, 30: 2830-2842.

[14] Li J, Wang C, Zhu H, et al. Efficient crowded scenes pose estimation and a new benchmark[EB/OL]. [2023-05-01]. https://arxiv.org/abs/1812.00324

[15] Moon G, Chang J Y, Lee K M. PoseFix: Model-agnostic general human pose refinement network[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019: 7773-7781.

[16] Zhou C, Chen S, Ding C, et al. Learning contextual and attentive information for brain tumor segmentation [C]//Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, Granada, Spain, 2019: 497-507.

[17] Ji X, Yang Q, Yang X, et al. Human pose estimation: multi-stage network based on HRNet[J]Journal of Physics, 2022, 2400(1): No. 012034.

[18] He K, Zhang X, Ren S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904-1916.

[19] Papandreou G, Zhu T, Kanazawa N, et al. Towards accurate multi-person pose estimation in the wild[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, USA, 2017: 4903-4911.

[20] Wei L, Zhang S, Dai J, et al. ST-GCN: spatial temporal graph convolutional networks for skeleton-based action recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Sait Lake City, USA, 2018: 7452-7461.

[21] Qin X, Zhang Z, Huang C, et al. U2-Net: going deeper with nested U-structure for salient object detection[J]. Pattern Recognition, 2020, 106: No. 107404.

[22] Zhou F, Zhu M, Bai J, et al. Deformable ConvNets v2: more deformable, better results[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Sait Lake City, USA, 2018: 9308-9316.

[23] Carreira J, Agrawal P, Fragkiadaki K, et al. Associative embedding: end-to-end learning for joint detection and grouping[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 2274-2284.

[24] Papandreou G, Zhu T, Chen L C, et al. PersonLab: person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model[C]//Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 2018: 282-299.

[25] Kreiss S, Bertoni A, Alahi A. PifPaf: composite fields for human pose estimation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 2019: 11977-11986.

[26] Insafutdinov M, Pishchulin L, Andres B, et al. DeepCut: joint subset partition and labeling for multi person pose estimation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: No. 533.

[27] Newell A, Yang A, Deng J. Stacked hourglass networks for human pose estimation[C]//Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, Holland, 2016 :483-499.

[28] Chen Y, Wang Z, Peng Y, et al. Cascaded pyramid network for multi-person pose estimation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Sait Lake City, USA, 2018: 7103-7112.

[29] Sun K, Xiao B, Liu D, et al. Deep high-resolution representation learning for human pose estimation [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, USA, 2019: 5693-5703.

[30] Xiao B, Wu H, Wei Y. Simple baseline for human pose estimation and tracking[C]//Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 2018: 742-757.

[31] Wei L, Zhang S, Yin W, et al. Convolutional pose machines[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Munich, Germany, 2016: 4724-4732.

[32] Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[33] Lin G, Li Q, Li M , et al. A novel bottleneck-activated feedback neural network model for time series prediction[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 32(4): 1621-1635.

[34] Cao Z, Hidalgo Martinez G, Simon T, et al. OpenPose: realtime multi-person 2D pose estimation using part affinity fields[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 43(1): 172-186.