文章编号：2096-1472(2023)-03-18-06
DOI:10.19644/j.cnki.issn2096-1472.2023.003.005

基于卷积与Transformer的人体姿态估计方法对比研究

冯 杰，郑建立

(上海理工大学健康科学与工程学院，上海 200093)
✉fjie666@outlook.com; zhengjianli163@163.com

摘 要：人体姿态估计是计算机视觉的基础性算法之一，为了探究人体姿态估计领域的研究发展趋势，文章首先介绍了基于卷积的经典人体姿态估计算法，论述各算法的基本原理及算法改进，其次对最新的基于自注意力模型(Transformer)的算法进行梳理，最后介绍了常用的公开数据集和模型评价指标，选取了几个经典算法进行对比分析，平均精度在马克斯·普朗克信息研究所(Max Planck Institute Informatik，MPII)数据集达到80%以上，在微软公共对象上下文(Common Objects in Context，COCO)数据集达到60%以上，得到卷积结构和Transformer结构互有优劣的结论。

关键词：姿态估计；关节点检测；卷积神经网络；Transformer
中图分类号：TP391.4 **文献标识码**：A

A Comparative Study of Human Pose Estimation based on Convolution and Transformer

FENG Jie, ZHENG Jianli

(*School of Health Science and Engineering*, *University of Shanghai for Science and Technology*, *Shanghai* 200093, *China*)
✉fjie666@outlook.com; zhengjianli163@163.com

Abstract: Human pose estimation is one of the basic algorithms in computer vision. In order to explore the research and development trend in the field of human pose estimation, this paper first introduces the classic human pose estimation algorithms based on convolution, and discusses the basic principles and algorithm improvements of each algorithm. Then, it reviews the latest algorithms based on the Self Attention Model (Transformer). Finally, it introduces the commonly used public datasets and model evaluation indicators. Several classical algorithms are selected for comparative analysis. The average accuracy is more than 80% in the dataset of Max Planck Institute Informatik (MPII), and more than 60% in the dataset of Microsoft Common Objects in Context (COCO). It is concluded that both convolution structure and Transformer structure have their own advantages and disadvantages.

Keywords: human pose estimation; joints detection; CNN (Convolutional Neural Networks); Transformer

1 引言(Introduction)

人体姿态估计(Human Pose Estimation，HPE)是计算机视觉中的一个重要任务，也是计算机理解人类动作和行为必不可少的一步。近年来，人体姿态估计正越来越多地应用于人们的日常生活，如在人机交互[1]和VR游戏领域对人体动作的捕捉[2]，在安防领域对人体行为的分析[3](如智能监控、肢体对抗等)，在运动和康复领域用于指导人的训练[4]。由于人体在执行部分动作时躯体姿态变化较大，以及动作背景环境复杂、观察角度的不确定，使人体姿态估计面临很多挑战，该领域正受到众多学者的密切关注。

自2012 年AlexNet[5]网络提出以来，深度学习得到蓬勃发展，给人体姿态估计领域带来了新的发展驱动力。2014 年，计算机视觉领域首次成功引入卷积神经网络解决单人姿态估计问题，在此后的很长一段时间内，基于卷积神经网络的骨干结构一直是该领域内的主流方法。随后，Transformer结构[6]在时序领域取得巨大成功，开始有研究者将其引入计算机视觉领域，基于Transformer结构的人体姿态估计算法成为新的研究热点。本文从卷积神经网络和基于Transformer结构的网络两个方面，对人体姿态估计算法做综合性论述，并总结分析了两种研究思路的优点和缺点。

2 人体姿态估计概述(Overview of human pose estimation)

人体姿态估计是指在视频或者图像中，对人体的关键点如肘部、手腕、膝盖等进行定位，并且能够计算得到各个关

基金项目：国家重点研发计划子课题(2020YFC2005802)。

节点之间的最优连接关系。单人姿态估计是指给定预测图像中只有单个人体或者固定数量的关节点。在深度学习被引入之前，传统处理姿态估计的方法常常是基于图结构模型[7]。图结构模型存在人工设计特征困难、鲁棒性低的问题，学者们发现基于深度学习不需要设计图模型的拓扑结构和关节点之间的交互，具有更大的优势。单人姿态估计可分为基于坐标回归的方法、基于热图检测的方法及混合模型方法。基于坐标回归和基于热图检测方法各有优劣，但由于基于坐标回归方法在精度上具有较大的局限性，因此目前主流方法仍然是基于热图检测。基于混合模型的方法，则是在一个算法中同时使用了前两者监督模型学习。表1中列出以上三种方法的优点和缺点对比。

表1 单人姿态估计算法的对比

Tab.1 Comparison of single-person estimation algorithms

方法	优点	缺点
基于坐标回归	获取坐标简单直观，计算量小，易于扩展到高维情况	高度非线性，模型学习困难且精度不足，鲁棒性不够
基于热图检测	精度高且具有良好的鲁棒性，适用于各种复杂场景	计算量大，算法运行效率低
混合模型	综合了坐标回归和热图检测的部分优点，能以较高精度得到坐标	结构复杂，训练复杂度高

多人姿态估计任务比单人姿态估计复杂，在图像中含有数量不等的多个人体。算法不仅需要给出所有关节点，还需要预测不同关节点分属的不同人体，即关节点分组的过程。目前，多人姿态估计主流方法为二步法，即必须经过两个阶段才能得到最终结果，二步法又分为自顶向下(Top-Down)和自底向上(Bottom-Up)两种方法。自顶向下的方法需要先在图像中检测人体，再在单个人体局部区域内做单人的姿态估计。自底向上的方法和自顶向下的方法相反，其过程是先将图像中所有关节点检测出来，然后使用分组算法将同一个人体的关节点连接起来。除二步法外，还有较为新颖的单步法。

自顶向下和自底向上方法各有优劣，自顶向下比较直观，但由于网络中还包含目标检测部分，因此运算效率不高。通常，需要高精度的场景，采用自顶向下的方法；对实时性要求比较高的场景，采用自底向上的方法。表2对两种方法的优劣进行对比。

表2 自顶向下和自底向上方法的优劣对比

Tab.2 Comparison of advantages and disadvantages of top-down and bottom-up methods

方法	优点	缺点
自顶向下	可改进人体检测器，关节定位精度高	需对图像中每个人进行人体检测，占用内存较多，效率低
自底向上	速度受图像人数的影响小，运行效率高	场景影响人体关节定位和分组，精度较低

3 基于卷积的算法(Algorithm based on convolution)

3.1 单人姿态估计

在单人姿态估计任务中，TOSHEV等[8]于2014 年首次将深度学习应用于人体姿态估计，并将其网络结构命名为DeepPose；其研究基于坐标回归的预测方法，从特征图中直接预测关键点的坐标，使用平方差损失函数进行回归计算损失值。DeepPose使用了一个级联回归预测，将训练分为多个阶段，以提高坐标回归的准确度。初始阶段得到粗略的坐标后，坐标点周围的局部图像被裁剪并送到下一个阶段的训练，学习更精细尺度的特征。这与目前流行的一些多尺度特征网络的思想有共通之处。

即使DeepPose已经使用级联回归进行预测，但让算法直接预测最终坐标值的做法对于模型来说仍然过于困难。这不仅是由于场景和人体动作的复杂多变，更是由于图像特征和关节坐标值之间是高度的非线性关系，是一个复杂的学习任务。之后，SZEGEDY等[9]在GoogleNet的基础上提出了误差迭代修正(Iterative Error Feedback，IEF)[10]方式改进此问题。误差迭代修正提出了通用型的修正回归误差方法，但是如何提高输出坐标的准确度，仍然没有行之有效的方法。TOMPSON等[11]较早地使用热图检测的方法进行姿态预测。研究者发现，相比较于坐标回归，基于热图检测的方法能够大幅度地提高算法对关节点的预测准确度。热图是由概率值代表的一副图像，图中像素点代表其为关节点的概率。此外，TOMPSON的研究贡献在于讨论了常规卷积神经网络中使用的池化层和Dropout会造成空间关联信息丢失，带来位置精度损失的问题，尤其是在姿态估计这种精细化任务中，特别需要这种特征信息。近年来，有越来越多的研究者关注和论证池化层会带来的信息丢失问题，不利于需要精确位置信息的任务。

之后，很多研究者大都从网络模型结构上进行精巧设计，如卷积姿态机[12]使用多个全卷积结构[13]网络预测关节的热图。NEWELL等[14]在2016 年提出Hourglass网络，其中的沙漏堆叠结构表现优秀，击败了以往所有的模型，成为一个经典的结构。Hourglass使用池化层和上采样构造沙漏形模块，使用残差结构将不同尺度特征进行融合，结合中间监督优化模型训练(图1)。

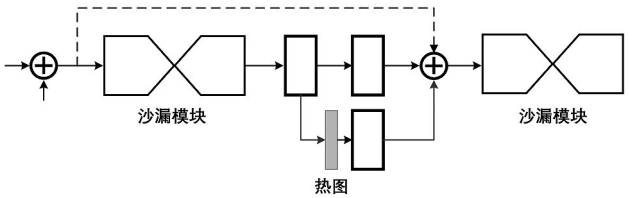

图1 中间监督

Fig.1 Intermediate supervision

基于Hourglass网络，其他研究者还提出了许多变种网络[15-16]，ZHANG等[17]对沙漏接口进行精简，提出轻量级沙漏网络，配合知识蒸馏降低模型复杂度,将知识从大型教师网络迁移到轻量级网络中。以上研究都基于一个思路，即设计复杂或者精巧的结构，期望用复杂结构进行姿态估计问题中的高度非线性拟合。XIAO等[18]提出简单基线网络，认为提高算法效果不一定要依赖复杂结构，XIAO的研究旨在提出一种简单的网络结构降低算法复杂度。简单基线网络如图2所示，算法通过常规顺序堆叠卷积层进行特征提取，使用反卷积进行分辨率的复原。

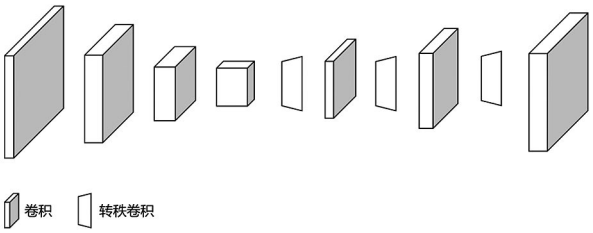

图2 简单基线网络

Fig.2 Simple baselines network

简单基线网络虽然网络结构简单，但是非常有效，能提示研究人员的网络学习能力已经饱和，另有影响算法表现的因素存在。2019年，微软团队提出高分辨率网络(High-Resolution Network，HRNet)[19]，研究认为不管使用池化层还是其他形式的图像下采样，降低分辨率的同时都会丢失特征，而高分辨率网络在基线上不需要降低分辨率，而是通过并行的子网分支下采样，通过不同尺度的感受野得到图像特征后，上采样叠加回基线分支进行交叉融合信息；其模型结构如图3所示，该结构目前仍然有优异的表现。

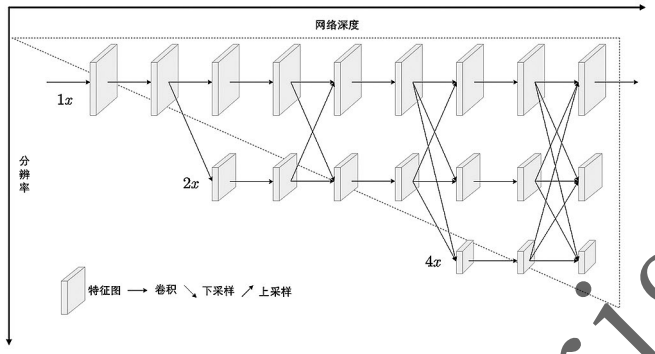

图3 高分辨率网络

Fig.3 High-resolution network

3.2 自顶向下方法

在多人姿态估计任务中，自顶向下方法对人体检测器依赖较大，需要准确得到单个人体局部图像。目前，大量的研究都集中在人体检测器上，针对多人姿态领域进行优化，希望得到高质量的检测框，其中对非极大值抑制的策略改进是众多论文的研究方向。FANG等[20]提出区域多人姿态估计框架(Regional Multi-person Pose Estimation，RMPE)，使用Faster R-CNN作为人体检测器，设计对称式变压器网络获取高精度的人体检测框，同时提出参数姿态非极大值抑制(P-Pose NMS)策略对冗余的检测框进行过滤，在检测框中配合沙漏堆叠网络进行单人姿态估计。谷歌团队将非极大值抑制与人体关节点评价指标关键点相似度(Object Keypoint Similarity，OKS)相结合，提出G-RMI[21]网络。不同于参数姿态非极大值抑制直接使用欧式距离进行过滤，OKP算法使用人体的尺度信息对临近的关节点间进行欧氏距离的修正，计算其检测框的相似度。同时，级联金字塔网络(Cascaded Pyramid Network，CPN)[22]算法也验证了不同非极大值抑制策略对于人体检测质量的影响。

3.3 自底向上方法

采用自底向上方法时，如何将所有关节点进行分组并联接得到人体拓扑结构是关键。CAO等[23]提出OpenPose网络是一种典型的自底向上的方法，OpenPose采用经典VGG-19作为主干网络提取特征，将特征输入到一个双分支网络，其中一个分支获取所有关节点热图，另一个分支获取部件亲和场(Part Affinity Fields，PAFs)，部件亲和场能将关节点进行分组和连接。PAPANDREOU等[24]提出多任务网络PersonLab，采用残差网络预测关节点热图，关节点偏移量及人体实例分割的掩模，利用基于树形运动学图的贪婪解码算法将关键点分组到人体检测实例中。

NEWELL等[25]提出关联嵌入标签算法，能够以端到端的方式对关节点进行检测和分组；其基本思想是为每次检测引入一个实数，用作识别对象所属组的“标签”，标签将每个检测与同一组中的其他检测相关联。NEWELL使用损失函数促使相同组的标签具有相似的值。

CHENG等[26]在高分辨率网络的基础之上，提出HigherHRNet，结合关联嵌入标签算法对关节点进行分组。NIE等[27]于2019年提出单阶段人体姿态器，它是一种新颖的单步法的多人姿态估计器，简化了人体估计的流程。本文提出了一种新的结构化关节的坐标表示方法，它首先使用根节点将人体进行基础的检测和定位，然后将关节点表示距离人体根节点的偏移。

以上经典的算法都基于卷积结构，同时有研究对热图损失进行分析。一般热图大小为原图的多倍下采样，从热图中取第一极大值并映射回原图坐标时，存在数学期望上的偏差。分布坐标感知(Distribution-Aware coordinate Representation of Keypoint，DarkPose)[28]和无偏数据处理(Unbiased Data Processing，UDP)[29]等算法对数据进行无偏处理，得到更精确的预测坐标，可无缝嵌入各种姿态估计模型中使用。

4 基于Transformer的算法(Algorithm based on Transformer)

Transformer是目前的热点研究方向。2020年，视觉自注意力模型(Vision Transformer，ViT)首次将Transformer结构引入计算机视觉领域。ViT将图像切分为$N\times N$大小的局部图像块作为序列，经过维度转换后传入Transformer模块，得到最终的输出特征。这种简单的切分图像作为序列输入的方式在小数据集上与同等规模的卷积神经网络相比并未取得最优秀的表现，但是在大数据集上的训练能得到出色的结果。这种结果是可预期的，Transformer缺乏卷积结构固有的平移不变性和局部特征性，因此当数据量不足时不能很好地拟合。针对这种原始Transformer参数量大和效果不佳的问题，有许多研究做出了改进。其中，移动窗口自注意力模型(Shift Windows Transformer，Swin-Transformer)[30]通过划分小窗口进行局部自注意力减少参数量，通过窗口滑动进行信息交换的方式，在各大任务中均超越卷积神经网络取得了顶尖的成绩。

使用Transformer进行人体姿态估计的研究目前不多，

其中姿态估计自注意力(Pose Estimation Transformer，PE-Former)与ViT结构相似，将图像切片后送入Transformer，但这种简单的设计使其效果并未达到领先水平。还有的研究将卷积结构与Transformer混合使用，例如直接自注意力估计算法(Transformer Pose，TFPose)[31]使用卷积神经网络作为骨干网络提取图像特征后，将特征添加位置嵌入输入Transformer模块，经过“编码—解码”结构的设计，得到最终的关节点输出。值得一提的是，TFPose并未使用常用的热图输出，而是直接对关节点坐标进行回归预测，其结构如图4所示。与TFPose相同，TransPose也是卷积与Transformer结合的网络设计，但是使用热图进行监督学习，使其效果优于TFPose。而高分辨率自注意力模型(High-Resolution Transformer，HRFormer)则基于高分辨率网络(High-Resololution Network，HRNet)骨干网络，将主体的卷积替换为Transformer结构，为了减少参数量，与Swin-Transformer类似，将特征图划分窗口进行自注意力学习，取得了不错的效果。

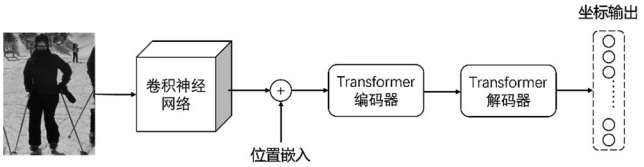

图4 TFPose网络

Fig.4 TFPose network

LI等[32]提出的基于级联Transformer的姿态识别(Pose Recognition with TRansformer，PRTR)研究构建了一个端到端可训练的自顶向下的多人姿态估计算法。该研究在网络内构建了人体检测器，并基于此人体检测器得到的检测框进行后续的关键点预测，算法中的人体检测器和关键点预测网络都是由Transformer构成的；而基于Transformer的自底向上的类型算法目前仍较少。

5 数据集与评价指标(Datasets and evaluation indicators)

目前，人体姿态估计领域内有许多公开的数据集，涵盖了单人的估计任务和多人的估计任务。其中，MPII数据集中既含有单人样本也包括多人样本；而像微软COCO竞赛数据集的样本数已经超过了30万张，是多人估计领域的一个重要数据集。表3和表4给出了常见的公开数据集。

表3 单人姿态估计数据集

Tab.3 Single-person pose estimation dataset

数据集	年份/年	关节点/个	样本数(约)/张	来源
LSP[33]	2010	14	2,000	Flickr下载的全身姿态图像
FLIC[34]	2013	10	20,000	从“好莱坞”电影中截取的视频帧
PennAction[35]	2013	13	2,000	YouTube的视频
JHMDB[36]	2013	15	1,000	来自动作识别数据集HMDB51
MPII[37]	2014	16	25,000	YouTube视频中截取的视频帧

表4 多人姿态估计数据集

Tab.4 Multi-person pose estimation dataset

数据集	年份/年	关节点/个	样本数(约)/张	来源
MPII[37]	2014	16	25,000	YouTube视频中截取的视频帧
COCO[38]	2014	17	330,000	谷歌、必应、Flickr下载的图片
HKD[39]	2017	14	300,000	互联网中的图片
PoseTrack[40]	2018	15	500	姿态跟踪数据集

对于如何评估算法的表现，常用的有4 个评估指标。①PCK：正确关键点的百分比。给定某一阈值，预测关节点与真实关节点的距离在阈值内的，视为正确。②PCP：正确部位百分比。两个预测关节点构成的肢体部位，与真实肢体关节距离在特定的阈值内的，视为正确。③PDJ：检测到的关节百分比。预测关节和真实关节之间的距离，在躯干直径某一百分比范围内。④OKS：对象关节点相似度。COCO关键点挑战竞赛采用这一评估指标。其中，OKS的计算公式见公式(1)：

$$OKS_p=\frac{\sum_i \exp\{-d_{pi}^2/2S_p^2\sigma_i^2\}\delta(v_{pi}>0)}{\sum_i \delta(v_{pi}>0)} \tag{1}$$

其中，p表示标注中某人；pi表示某人的其中一个关键点；d_{pi}^2表示第p 个人第i 个关键点检测位置与标注位置坐标的欧式距离平方$d_{pi}^2=\left(x_i^{'}-x_{pi}\right)^2+\left(y_i^{'}-y_{pi}\right)^2$；$v_{pi}=1$表示这个关键点无遮挡且已标注，$v_{pi}=2$表示这个关键点有遮挡但已标注；$\delta$函数表示当关键点被标注时才纳入计算。$S_p$表示标注的行人的尺度因子，$S_p=\sqrt{wh}$，$w$、$h$为检测框的宽、高；$\sigma_i$为第$i$ 个关键点的类型对应的归一化因子，可以视为一个常数。

6 实验结果对比(Comparison of experimental results)

表5给出一些单人姿态估计算法在MPII数据集上的实验结果对比，以0.5阈值的PCK为评估指标，计算所有类型关节点的平均精度。表6给出一些多人姿态估计算法在COCO数据集上的实验结果对比，以OKS为评价指标，计算所有类型关节点的平均精度。

表5 单人姿态估计算法在MPII数据集上的表现

Tab.5 Results of single-person pose estimation algorithm on dataset MPII

算法	年份/年	平均精度/%
IEF	2015	81.3
CPM	2016	88.5
Hourglass	2016	90.9
HRUs	2017	91.5
SGANPose	2018	91.8
Simple baselines	2018	91.5
FPD	2019	91.1
HRNet	2019	92.3
Adversarial semantic	2020	94.1
TFPose	2021	90.4
PRTR	2021	89.5

表6 多人姿态估计算法在COCO数据集上的表现

Tab.6 Results of multi-person pose estimation algorithm on dataset COCO

算法	方法	年份/年	平均精度/%
RMPE	自顶向下	2017	61.8
G-RMI	自顶向下	2017	64.9
CPN	自顶向下	2018	73
HRNet	自顶向下	2019	75.8
Adversarial semantic	自顶向下	2020	75.2
PRTR	自顶向下	2021	73.3
TransPose	自顶向下	2021	75.3
HRFormer	自顶向下	2021	75.6
OpenPose	自底向上	2017	61.8
Associative Embedding	自底向上	2017	65.5
PersonLab	自底向上	2018	68.7
HigherHRNet	自底向上	2019	70.5
直接姿态估计(DirectPose)	自底向上	2021	74.8

从表5中可以看出，沙漏(Hourglass)网络凭借其独特的结构在算法表现上取得了较大的突破，MPII数据集中的平均精确度突破90%。之后其他研究中的网络结构大体上保留“下采样—上采样”的沙漏形的设计痕迹，如简单基线网络整体上可视为一个沙漏形。这种两头大中间小的模型设计，在卷积神经网络的维度设计中也运用广泛，称之为瓶颈(Bottleneck)结构，其特点是首先通过1×1卷积降低维度，然后进行常规的$N\times N$卷积，再使用1×1卷积将维度升高还原。近年来，出现了逆瓶颈层的设计，通过先升高维度提取更多特征后再降低维度。逆瓶颈层的结构在姿态估计中是否能起到效果，是一个值得探讨的问题。HRNet相比其他算法，表现更优异，这在很大程度归因于其网络全程保持与热图一致的高分辨率，也证明了特征图的分辨率对预测结果具有较大影响。

人体姿态估计自顶向下的方法优于自底向上方法在前文已做介绍，从表6中可以看出自底向上方法指标与自顶向下方法的指标相比仍有较大差距，其主要原因是自顶向下的方法经过检测器后得到单个人体图像，可以视为带有先验知识，即局部图像中的人体关节具有某种拓扑连接规律，如头部之下为肩颈等。这种全体样本都具备的特征规律能够很好地指导算法学习，得到准确的关节点。自底向上方法由于需要先检测图像中所有关节点，在图像人体数量众多的情况下丧失了这种先验知识，加之关节点分布凌乱，导致误检率、漏检率较高。如何将人体拓扑结构这种先验知识带入自底向上的方法，也是一个值得研究的方向。

Transformer在姿态估计中的应用仍然是一个新的研究方向，从表6中可以看出，基于Transformer方法的指标表现与基于卷积方法的指示表现大致持平。HRFormer在HRNet的基础上，将卷积替换成Transformer结构后，仅带来准确率的微小提升。Transformer本身是在时序领域提出的，虽然目前在图像分类领域成为最先进的结构，但是在特定视觉任务姿态估计中，语义分割等未取得突破性的提升。视觉任务的特征本身在空间域的相关性较高，简单地将其空间展开后模拟成时间域并不能很好地捕捉其特征关系，梳理与处理这两者间的转化，或许能成为Transformer提升姿态估计表现的关键。近期，有研究开始回归卷积神经网络本身，Facebook的存粹卷积模型(ConvNeXt)仅凭借卷积结构和其他算法的设计细节结合，便在大规模视觉识别挑战赛(Large Scale Visual Recognition Challenge，ILSVRC)图像分类数据集上达到了目前最好的Top-1的准确率。基于此，卷积与Transformer，谁更有潜力，开始成为研究者讨论的热点。

7 结论(Conclusion)

综上所述，人体姿态估计领域依托于深度学习的发展，展现出了巨大的潜力和优异的表现。目前，基于卷积结构的算法具有简单、高效的优点，仍是该领域最具竞争力的算法，基于Transformer结构的新颖算法有着巨大的发展潜力。算法精度与执行速度兼顾的平衡将会是该领域的研究重点，未来随着深度学习基础性理论的发展，将会诞生更高效的模型和研究成果。

参考文献(References)

[1] 唐彪,樊启润,孙开鑫(等).人体姿态识别算法在视觉人机交互中的应用[J].计算机测量与控制,2019,27(7):242-247.

[2] 张继凯,顾兰君.基于骨架信息的人体动作识别与实时交互技术[J].内蒙古科技大学学报,2020,39(3):266-272.

[3] 马子健,林雨衡,王志强,等.封闭环境下人体姿态识别及打架行为监测[J].计算机应用,2021,41(S2):214-220.

[4] 闫航,陈刚,佟瑶,等.基于姿态估计与GRU网络的人体康复动作识别[J].计算机工程,2021,47(1):12-20.

[5] KRIZHEVSKY A, SUTSKEVER I, HINTION G E. ImageNet classification with deep convolutional neural networks[J]. Advances in Neural Information Processing Systems, 2012, 25:1097-1105.

[6] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30:5998-6008.

[7] FELZENSZWALB P F, HUTTENLOCHER D P. Pictorial structures for object recognition[J]. International Journal of Computer Vision, 2005, 61(1):55-79.

[8] TOSHEV A, SZEGEDY C. DeepPose: human pose estimation via deep neural networks[C]// Institute of Electrical and Electronics Engineers. Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Piscataway: IEEE Computer Society, 2014: 1653-1660.

[9] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions[C]// Institute of Electrical and Electronics Engineers. Proceedings of the 2015IEEE Conference on Computer Vision and Pattern Recognition, Piscataway: IEEE Computer Society, 2015:1-9.

[10] CARREIRA J, AGRAWAL P, FRAGKIADAKI K, et al. Human pose estimation with iterative error feedback[C]// Institute of Electrical and Electronics Engineers. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Piscataway: IEEE Computer Society, 2016: 4733-4742.

[11] TOMPSON J, GOROSHIN R, JAIN A, et al. Efficient object localization using convolutional networks[C]// Institute of Electrical and Electronics Engineers.Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition, Piscataway: IEEE Computer Society, 2015: 648−656.

[12] WEI S E, RAMAKRISHNA V, KANADE T, et al. Convolutional pose machines[C]// Institute of Electrical and Electronics Engineers. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Piscataway: IEEE Computer Society, 2016: 4724−4732.

[13] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]// Institute of Electrical and Electronics Engineers. Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition, Piscataway: IEEE Computer Society, 2015: 3431−3440.

[14] NEWELL A, YANG K, DENG J. Stacked hourglass networks for human pose estimation[C]// Springer Science. Proceedings of the 14th European Conference on Computer Vision, Berlin: Springer, 2016:483−499.

[15] CHU X, YANG W, OUYANG W, et al. Multi−context attention for human pose estimation[C]// Institute of Electrical and Electronics Engineers. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Piscataway: IEEE Computer Society, 2017:5669−5678.

[16] BIN Y, CAO X, CHEN X, et al. Adversarial semantic data augmentation for human pose estimation[C]// Springer Science. Proceedings of the 16th European Conference on Computer Vision, Berlin: Springer, 2020:606−622.

[17] ZHANG F, ZHU X, YE M. Fast human pose estimation[C]// Institute of Electrical and Electronics Engineers. Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition, Piscataway: IEEE Computer Society, 2019:3517−3526.

[18] XIAO B, WU H, WEI Y. Simple baselines for human pose estimation and tracking[C]// Springer Science. Proceedings of the 15th European Conference on Computer Vision, Berlin: Springer, 2018:472−487.

[19] SUN K, XIAO B, LIU D, et al. Deep high−resolution representation learning for human pose estimation[C]// Institute of Electrical and Electronics Engineers. Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition, Piscataway: IEEE Computer Society, 2019:5693−5703.

[20] FANG H S, XIE S, TAI Y W, et al. RMPE: regional multi−person pose estimation[C]// Institute of Electrical and Electronics Engineers. Proceedings of the 2017 IEEE International Conference on Computer Vision, Piscataway: IEEE Computer Society, 2017:2353−2362.

[21] PAPANDREOU G, ZHU T, KANAZAWA N, et al. Towards accurate multi−person pose estimation in the wild[C]// Institute of Electrical and Electronics Engineers. Proceedings of the 2017 IEEE International Conference on Computer Vision, Piscataway: IEEE Computer Society, 2017:4903−4911.

[22] CHEN Y, WANG Z, PENG Y, et al. Cascaded pyramid network for multi−person pose estimation[C]// Institute of Electrical and Electronics Engineers. Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, Piscataway: IEEE Computer Society, 2018:7103−7112.

[23] CAO Z, SIMON T, WEI S E, et al. Realtime multi−person 2D pose estimation using part affinity fields[C]// Institute of Electrical and Electronics Engineers. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Piscataway: IEEE Computer Society, 2017:1302−1310.

[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]// Springer Science. Proceedings of the 15th European Conference on Computer Vision, Berlin: Springer, 2018: 282−299.

[25] NEWELL A, HUANG Z, DENG J. Associative embedding: end−to−end learning for joint detection and grouping[J]. Advances in Neural Information Processing Systems, 2017, 30:2277−2287.

[26] CHENG B, XIAO B, WANG J, et al. Higherhrnet: Scale−aware representation learning for bottom−up human pose estimation[C]// Institute of Electrical and Electronics Engineers. Proceedings of the 2020 IEEE Conference on Computer Vision and Pattern Recognition, Piscataway: IEEE Computer Society, 2020:5386−5395.

[27] NIE X, FENG J, ZHANG J, et al. Single−stage multi−person pose machines[C]// Institute of Electrical and Electronics Engineers. Proceedings of the 2019 IEEE International Conference on Computer Vision, Piscataway: IEEE Computer Society, 2019:6951−6960.

[28] ZHANG F, ZHU X, DAI H, et al. Distribution−aware coordinate representation for human pose estimation[C]// Institute of Electrical and Electronics Engineers. Proceedings of the 2020 IEEE Conference on Computer Vision and Pattern Recognition, Piscataway: IEEE Computer Society, 2020:7093−7102.

[29] HUANG J, ZHU Z, GUO F, et al. The devil is in the details: delving into unbiased data processing for human pose estimation[C]// Institute of Electrical and Electronics

Engineers. Proceedings of the 2020 IEEE Conference on Computer Vision and Pattern Recognition, Piscataway: IEEE Computer Society, 2020:5700-5709.

[30] LIU Z, LIN Y, CAO Y, et al. Swin Transformer: hierarchical vision transformer using shifted windows[C]// Institute of Electrical and Electronics Engineers. Proceedings of the 2021 IEEE International Conference on Computer Vision, Piscataway: IEEE Computer Society, 2021:10012-10022.

[31] YANG S, QUAN Z, NIE M, et al. TransPose: keypoint localization via transformer[C]// Institute of Electrical and Electronics Engineers. Proceedings of the 2021 IEEE International Conference on Computer Vision, Piscataway: IEEE Computer Society, 2021:11802-11812.

[32] LI K, WANG S, ZHANG X, et al. Pose recognition with cascade transformers[C]// Institute of Electrical and Electronics Engineers. Proceedings of the 2021 IEEE International Conference on Computer Vision, Piscataway: IEEE Computer Society, 2021:1944-1953.

[33] JOHNSON S, EVERINGHAM M. Clustered pose and nonlinear appearance models for human pose estimation[C]// British Machine Vision Association. Proceedings of the 2010 British Machine Vision Conference, London:British Machine Vision Association, 2010:1-11.

[34] SAPP B, TASKAR B. MODEC: multimodal decomposable models for human pose estimation[C]// Institute of Electrical and Electronics Engineers. Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, Piscataway: IEEE Computer Society, 2013:3674-3681.

[35] ZHANG W, ZHU M, DERPANIS K G. From actemes to action: A strongly-supervised representation for detailed action understanding[C]// Institute of Electrical and Electronics Engineers. Proceedings of the 2013 IEEE International Conference on Computer Vision, Piscataway: IEEE Computer Society, 2013:2248-2255.

[36] JHUANG H, GALL J, ZUFFI S, et al. Towards understanding action recognition[C]// Institute of Electrical and Electronics Engineers. Proceedings of the 2013 IEEE International Conference on Computer Vision, Piscataway: IEEE Computer Society, 2013:3192-3199.

[37] ANDRILUKA M, PISHCHULIN L, GEHLER P, et al. 2D human pose estimation: new benchmark and state of the art analysis[C]// Institute of Electrical and Electronics Engineers. Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Piscataway: IEEE Computer Society, 2014:3686-3693.

[38] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft coco: common objects in context[C]// Springer Science. Proceedings of the 13th European Conference on Computer Vision, Berlin: Springer, 2014:740-755.

[39] WU J, ZHENG H, ZHAO B, et al. Large-scale datasets for going deeper in image understanding[C]// Institute of Electrical and Electronics Engineers. Proceedings of the 2019 IEEE International Conference on Multimedia and Expo (ICME), Piscataway: IEEE Computer Society, 2019: 1480-1485.

[40] ANDRILUKA M, IQBAL U, INSAFUTDINOV E, et al. PoseTrack: a benchmark for human pose estimation and tracking[C]// Institute of Electrical and Electronics Engineers. Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, Piscataway: IEEE Computer Society, 2018:5167-5176.

作者简介：

冯　杰(1994-)，男，硕士生.研究领域：计算机视觉.

郑建立(1965-)，男，博士，副教授.研究领域：医学信息集成.本文通信作者.

(上接第45页)

mbmodel size[C]// ICLR Organizing Committee. ICLR' 17 Conference Proceedings. Toulon: International Conference on Learning Representations, 2017:207-212.

[7] HOWARD A, ZHU M, CHEN B, et al. MobileNets: efficient convolutional neural networks for mobile vision applications[C]// CVPR Organizing Committee. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hawaii: IEEE Computer Society, 2017: 1704-1712.

[8] ZHANG X, ZHOU X, LIN M, et al. ShuffleNet: An extremely efficient convolutional neural network for mobile devices[C]// CVPR Organizing Committee. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE Computer Society, 2018:6848-6856.

[9] 董子源,韩卫光.基于卷积神经网络的垃圾图像分类算法[J].计算机系统应用,2020,29(8):199-204.

[10] 徐传运,王影,王文敏,等.面向生活垃圾图像分类的多级特征加权融合算法[J].重庆理工大学学报(自然科学),2022,36(09):146-155.

[11] SANDLER M, HOWARD A, ZHU M, et al. MobileNetV2: Inverted residuals and linear bottlenecks[C]// CVPR Organizing Committee. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE Computer Society, 2018:4510-4520.

作者简介：

易才键(1998-)，男，硕士生.研究领域：深度学习，计算机视觉.

陈　俊(1978-)，男，硕士，副教授.研究领域：物联网通信.

王师玮(1998-)，女，硕士生.研究领域：计算机视觉.