计算机科学
COMPUTER SCIENCE

基于局部特征与全局表征耦合的2D人体姿态估计

陈乔松, 吴济良, 蒋波, 谭冲冲, 孙开伟, 邓欣, 王进

引用本文

陈乔松, 吴济良, 蒋波, 谭冲冲, 孙开伟, 邓欣, 王进. 基于局部特征与全局表征耦合的2D人体姿态估计[J]. 计算机科学, 2023, 50(11A): 221100007-5.

CHEN Qiaosong, WU Jiliang, JIANG Bo, TAN Chongchong, SUN Kaiwei, DEN Xin, WANG Jin. Coupling Local Features and Global Representations for 2D Human Pose Estimation [J]. Computer Science, 2023, 50(11A): 221100007-5.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.221100007

基于局部特征与全局表征耦合的 2D 人体姿态估计

陈乔松[1]　吴济良[1]　蒋　波[1]　谭冲冲[2]　孙开伟[1]　邓　欣[1]　王　进[1]
1 重庆邮电大学计算机科学与技术学院　重庆 400065
2 重庆邮电大学自动化学院/工业互联网学院　重庆 400065

摘　要　近年来卷积神经网络和 Transformer 都在人体姿态估计领域中取得进步，卷积神经网络(Convolutional neural network，CNN)擅长提取局部特征，Transformer 擅长捕捉全局表征，但目前结合两者实现人体姿态估计的研究较少且效果不佳。针对此问题，提出一种耦合局部特征和全局表征的的模型 CNPose(CNN-Nest Pose)，该框架的局部-全局特征耦合模块利用多头注意力计算和残差结构的方式深度耦合局部特征和全局表征；还提出了局部-全局信息交流模块解决局部-全局特征耦合模块在计算过程中局部特征和全局表征数据源范围不一致的问题。在 COCO-val2017 和 COCO-dev-test2017 数据集上进行了验证，实验表明，采用了局部特征和全局表征耦合的 CNPose 模型相较于同类型方法有着更为优越的表现。

关键词：人体姿态估计；Transformer；卷积神经网络；局部特征；全局表征；特征耦合；注意力
中图法分类号　TP391.4

Coupling Local Features and Global Representations for 2D Human Pose Estimation

CHEN Qiaosong[1]，WU Jiliang[1]，JIANG Bo[1]，TAN Chongchong[2]，SUN Kaiwei[1]，DEN Xin[1] and WANG Jin[1]
1 School of Computer Science and Technology，Chongqing University of Posts and Telecommunications，Chongqing 400065，China
2 School of Automation/School of Industrial Internet，Chongqing University of Posts and Telecommunications，Chongqing 400065，China

Abstract　In recent years，both convolutional neural network and Transformer have made progress in the field of human pose estimation. Convolutional neural network(CNN) is good at extracting local features，and Transformer does well in capturing global representations. However，there are few studies on the combination of the two to achieve human pose estimation，as the same time the results are not good. Aiming at solving this problem，this paper proposes a model CNPose(CNN-Nest Pose) that couples local features and global representations. The local-global feature coupling module of this framework uses multi-head attention calculation method and residual structure to deeply couple local features and global representations. At the same time this paper proposes a local-global information exchange module to solve the problem that the range of data sources of local features and global representation is inconsistent in the local-global feature coupling module during the calculation process. The CNPose framework has been verified on COCO-val2017 and COCO-dev-test2017 datasets. Experiment results show that the CNPose model using the coupling of local features and global representations has superior performance compared to similar methods.

Keywords　Human pose estimation，Transformer，Convolutional neural networks，Local features，Global representations，Feature coupling，Attention

1　引言

人体姿态估计(Human Pose Estimation)是计算机视觉中十分具有挑战性的任务之一，其旨在让机器识别出输入图像中人体关节点的空间位置并生成一副能代表姿态的人体骨骼框架，其实现过程主要有两种方式：(1)自顶向下(Top-down)[1-2]；(2)自底向上(Bottom-up)[3-4]。自顶向下的方法先检测图像中每个人体的位置并用边界框标定，然后在每个边界框内计算人体关节点位置；与之相反，自底向上的方法先计算出图中所有关节点位置，然后利用人体模型拟合或者相关算法进行分组形成独立的人体骨骼框架。

2012 年 AlexNet[5] 在 ImageNet 图像分类挑战赛上取得冠军，这开启了深度卷积神经网络在计算机视觉领域的元年。在之后的发展过程中，研究者们设计出了一系列十分优秀的基础特征提取网络，如 ResNet[6]，VGG[7]，MobileNet[8] 等。2014 年 Alexander 等[9]较早将深度学习引入人体姿态估计领域，摆脱了手工设计的繁琐步骤，随后研究者们又设计出精度更高的网络，如 Pose Machines[10]，HRNet[11] 等，这些网络都在 COCO 等数据集上取得了前沿效果。上述网络都使用多尺度方法构建特征金字塔的方式来提取更为丰富的特征。卷积神经网络的卷积计算主要是计算图像中相邻像素或者是特征图中相邻部分之间的关系，同时卷积计算也是一种模板计算，特征图的不同部位的计算权重都相同，因此卷积神经网络提取特征过程具有局部性。

与卷积神经网络不同的是，Transformer 可以捕捉整个特征图中任意两部分之间的长短距离依赖，因此 Transformer

基金项目：国家重点研发项目(2022YFE0101000)
This work was supported by the National Key Research and Development Program of China(2022YFE0101000).
通信作者：陈乔松(chenqs@cqupt.edu.cn)

的注意力计算过程具有全局性。2020 年 Mao 等[12]提出了一种结合卷积神经网络和 Transformer 的串行结构网络 TFPose 用于人体姿态估计，该网络先使用卷积神经网络提取局部特征，再利用 Transformer 结构在局部特征中捕捉长短距离关系，这种结合方式较为简单。

综上可知，由于卷积计算是一种局部计算，因此卷积神经网络提取的特征具有局部性（下文将该特征简称为局部特征）；注意力计算是一种全局性计算，因此 Transformer 捕捉的长短距离关系具有全局性（下文将其简称为全局表征）。这两种特征对人体姿态估计都十分重要，但是前人利用这两种特征实现人体姿态估计的结合程度都比较低。

针对上述问题，本文基于卷积神经网络和 Transformer 提出了一种深度耦合局部特征和全局表征的串行网络 CNPose。

（1）CNPose 网络的核心模块是局部-全局特征耦合模块，该模块采用注意力计算和残差的方式耦合局部特征和全局表征，使得局部特征包含了更多的全局信息，全局特征包含了丰富的细节信息。

（2）为了解决局部-全局特征耦合模块在注意力计算时两种特征数据源范围差异带来的精度丢失的问题，本文提出了局部-全局交流模块来弥补上述数据源范围差异，同时该结构丰富了块与块之间的注意力。

2 相关工作

2.1 轻量级卷积结构 Bottleneck

卷积神经网络提取特征时，如果特征图的通道数目越多，卷积核参数也会越大，这会增加卷积神经网络的参数量和计算量。轻量级卷积结构 Bottleneck 能有效解决这个问题。Bottleneck 结构中有 3 个卷积层，首尾卷积层的卷积核大小都是 1×1，主要用于控制特征图通道数目，中间卷积结构的卷积核大小一般是 3×3，用于提取特征。在计算时，先利用头部的卷积层将特征图的通道数目减小，中间卷积层在与规模减小的特征图计算时需要的卷积核参数相应减少，最后一层卷积层用于恢复特征图的通道数目。Bottleneck 的设计实现了在不改变特征图的通道数目的情况下减少了参数量，因此本文也采用 Bottleneck 结构提取局部特征。

2.2 Nest：聚合嵌套的 Transformer

2021 年 Zhang 等[13]提出了聚合嵌套结构的 Nest Transformer。Nest Transformer 将图片分成不同的块，每一个块再次均分并编码成 Patch Token。这个过程满足式（1）：

$$T_n \times n = \frac{H \times W}{S^2} \tag{1}$$

其中，T_n 代表初次划分块的数量，n 代表每个 Patch Token 的维度，H 和 W 分别代表图片的高度和宽度，S 代表块均分编码时的大小。Nest Transformer 在计算过程中自注意力范围局限在每一个块内，它的亮点是块与块之间的聚集嵌入，每 4 个块聚集成一个块，聚集后的块包含的语义信息维度更高，聚集后块数量减小到聚集前的 25%，块中 Patch Token 的维度增加到原来的 2 倍。

3 基于局部特征与全局表征耦合的 2D 人体姿态估计

本节分为 3 部分：首先介绍 CNPose 的整体结构；其次为了高效地耦合卷积神经网络提取的局部特征和 Transformer 捕捉的全局表征，本文设计了一个局部-全局特征耦合模块；最后，卷积神经网络在提取特征的过程中是局部连续的，而 Nest Transformer 的多头自注意力计算被限制在互不重叠的区域中，为了减小局部-全局特征耦合模块计算过程中两种特征的数据源范围差异，同时实现更为精细的特征耦合，本文在局部-全局特征耦合模块的基础上设计了基于特征混淆的局部-全局信息交流模块。

3.1 CNPose 框架

为了耦合卷积神经网络提取特征的局部性和 Transformer 提取特征的全局性，本文提出了如图 1 所示的网络结构。

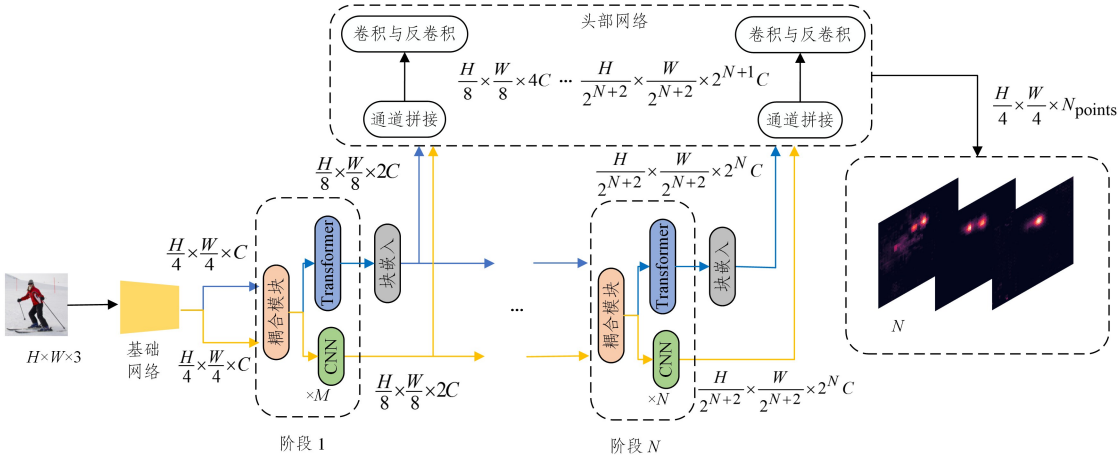

图 1 CNPose 结构图（电子版为彩图）

Fig. 1 CNPose structure diagram

CNPose 框架是一个串行结构，其计算分为多个阶段，前一阶段的输出为后一阶段的输入，同时每个阶段的输出都会进入头部网络并成为整体输出的一部分。图中蓝色线表示 Nest Transformer 的计算路径，橙色线表示卷积神经网络的计算路径，黑色线表示卷积神经网络和 Nest Transformer 共享的计算路径。在每个阶段中，局部特征和全局表征首先会经过局部-全局特征耦合模块，使得局部特征中包含全局信息，全局表征中包含局部信息。经过耦合后的局部特征进入 Bottleneck，耦合后全局表征进入 Nest Transformer 模块。每个阶段结束时全局特征图进行块聚集嵌入计算，得到的特征

图维度减小，可以表示距离更远的长距离依赖关系。

3.2 局部-全局特征耦合模块

卷积神经网络主要是计算相邻像素或特征图中相邻部分之间的关系，因此卷积神经网络计算具有局部性；Transformer 计算图像或者特征图中任意两部分之间的相互依赖关系，两部分之间可以没有距离与范围的限制，因此 Transformer 计算的结果具有全局性。为了让网络不仅能提取范围更广的全局表征，还能提取更为丰富的局部细节信息，本文提出局部-全局特征耦合模块，旨在深度耦合卷积神经网络提取的局部特征和 Transformer 提取的全局特征，增强模型的特征提取能力。局部-全局特征耦合模块详细结构如图 2 所示。

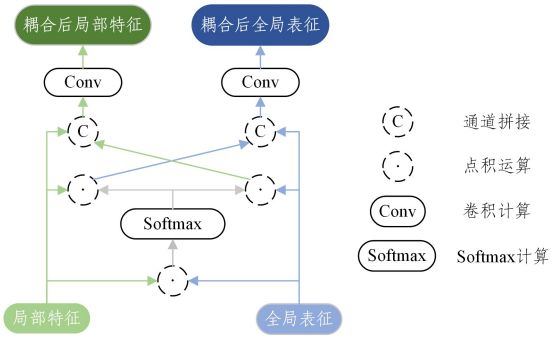

图 2 局部-全局特征耦合模块

Fig. 2 Local-Global feature coupling module

局部-全局特征耦合模块是一个多头注意力计算的残差结构。CNPose 分为多个计算阶段，上一阶段得到的局部特征和全局表征经过点积运算后得到相似矩阵。为了丰富局部特征的全局信息，相似矩阵经过 softmax 计算后与全局表征进行点积计算。为了保证局部特征的性质。上一步计算的结果与局部特征进行通道拼接并使用卷积进行维度调整，得到的结果便是耦合后的局部特征。丰富全局表征的局部性过程与此类似。局部-全局特征耦合模块的计算过程如式(2)、式(3)所示：

$$\boldsymbol{l}_c = Conv\left(\boldsymbol{g} \cdot \mathrm{Softmax}\left(\frac{\boldsymbol{lg}}{\sqrt{d_l}}\right) + \boldsymbol{l}\right) \tag{2}$$

$$\boldsymbol{g}_c = Conv\left(\boldsymbol{l} \cdot \mathrm{Softmax}\left(\frac{\boldsymbol{gl}}{\sqrt{d_g}}\right) + \boldsymbol{g}\right) \tag{3}$$

在式(2)和式(3)中，$\boldsymbol{l}_c$ 和 $\boldsymbol{g}_c$ 分别代表耦合后的局部特征和全局表征，$Conv$ 代表卷积运算，$\boldsymbol{l}$ 和 $\boldsymbol{g}$ 分别代表耦合前的局部特征和全局表征，$\sqrt{d_l}$ 和 $\sqrt{d_g}$ 代表 $\boldsymbol{l}$ 和 $\boldsymbol{g}$ 向量的维度。

使用注意力计算的方式能有效耦合局部特征和全局特征的原因是：人体关节点具有对称性，比如左手腕与右手腕、左膝盖与右膝盖等，经过卷积神经网络计算后，局部特征中对称关节点周围的分布更为接近，经过 Nest Transformer 的多头自注意力计算后，全局表征表示原特征图中区域间的相对关联程度，显然关节点尤其是对称关节点之间的关联程度更高，其中关联程度通过 Loss 函数进行评判，并且通过反向传播过程不断地调整。除此之外，点积运算的结果可以用来评判参与计算的两个向量的相似程度，值越大代表相似程度越高，因此在使用注意力计算的方式耦合局部特征和全局表征的过程中，局部特征矩阵和全局表征矩阵经过点积计算得到相似矩阵，相似矩阵中含有对称关节点区域的值更大。相较于单独使用局部特征或者全局特征做自注意力计算过程中得到的相似矩阵，局部特征和全局表征的注意力计算过程中的相似矩阵更为精细准确。卷积神经网络善于提取纹理等细节信息，对称关节点周围的纹理信息更为相似，Nest Transformer 擅长捕捉长短距离依赖关系，形成关节点之间的空间相对关系。使用局部-全局特征耦合模块耦合两种特征后，网络既能提取丰富的细节信息，还能捕捉更为精确的长短距离关系。当图像中的人体关节点因为服饰的变化导致对称关节点周围的纹理信息差异较大时，网络能够通过提取的空间相对关系还原出这些难点。

3.3 局部-全局信息交流模块

在局部-全局特征耦合模块中，注意力计算范围被限制在不重叠的块中，但是卷积神经网络计算过程是连续且重叠的，因此在耦合计算过程中全局表征的计算来源要小于局部特征的计算来源，即局部特征和全局表征的数据源范围存在差异，这会影响模型的精度。Transformer 的全局性是局限在块范围内，在块聚集嵌入前块与块之间并无关联，但是在图像中块与块之间存在较大的相似性，比如位于不同块中的左右手腕周围特征，但在块聚集融合之后，块对应原图表示的范围更大，虽然可以计算更大范围内的注意力，但是无法做到精细化计算上述的左右手腕两个局部相似区域的注意力。为了解决上述问题，本文提出了局部-全局信息交流模块，结构如图 3 所示。

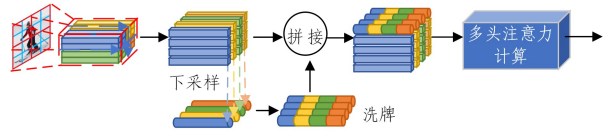

图 3 局部-全局信息交流模块

Fig. 3 Local-Global communication module

图像或者特征图被划分成不重叠的块，在块内再次被细分后编码成不同的 Patch Token。块内的所有 Patch Token 经过下采样得到局部-全局交流令牌(图中的圆柱形表示的特征)，局部-全局交流令牌代表此块的高层语义信息。对局部-全局交流令牌等分洗牌后，得到的每一个新的局部-全局交流令牌中都包含所有块的高层语义信息，再将新的局部-全局交流令牌与原有分块的 Patch Token 进行通道拼接。经过上述的计算后，每一个块中不仅包含了代表自身语义的信息，还包含了其他块少量的高层语义信息。

上述计算过程使得局部特征和全局表征的分块中都包含其他块的语义信息，这个过程减小了局部特征和全局表征的数据源范围的差异，这样的好处是得到的相似矩阵更为精确，能更有效耦合局部特征和全局表征。数据源范围的差异减小后，在注意力计算过程中，不仅可以得到块内不同部分的相互注意力，还可以得到块内 Patch Token 与其他块之间的较弱注意力。Transformer 的块聚集嵌入计算后无法做到精细化计算两个局部相似区域的注意力，但经过局部-全局信息交流模块得到的块内 Patch Token 与其他块之间的较弱注意力可以弥补这种不足。

4 实验

4.1 数据集

本文采用 COCO2017[14] 数据集来训练并验证模型，数据集中实例主要是中大尺寸。本实验使用 COCO2017 train 训练集作为训练集，其中包含 118 287 张图片和 149 813 个可见

的人体实例，实验的验证集为 COCO2017 val 数据集，其中包含 5000 张图片和 6352 个可见的人体实例。

4.2 评价指标

本文采用 OKS(Object Keypoint Similarity)[15]指标来评估模型定位人体关节点的准确性，计算过程如式(4)所示：

$$OKS_p=\frac{\sum_i \exp\left\{\frac{-d_{pi}^2}{2S_p^2\sigma_i^2}\right\}\delta(v_{pi}=1)}{\sum_i \delta(v_{pi}=1)} \tag{4}$$

其中，p 表示数据集中人的 ID；i 表示人体关节点的 ID；d_{pi} 表示对于 ID 为 p 的人的 ID 为 i 的关节点，模型预测的位置和真实的位置的欧氏距离；S_p 表示 ID 为 p 的人在标记数据中所占面积比例；σ_i 表示 ID 为 i 的关节点的归一化因子，数值由数据集统计得到，值越大表示该关节点越难标注；v_{pi} 表示 ID 为 p 的人的 ID 为 i 的关节点是否可见。δ 公式表示真实关节点可见时才纳入计算。

4.3 实验细节

本文实验环境为 Ubuntu 18.04 系统，CPU 为 Intel(R) Core(TM) i5-8400，内存为 64 GB，显卡为两张 NVIDIA V100，共 64GHz 显存。论文采用自顶向下的技术方案，实验过程中，数据集中的人体实例被切割成单人实例，并将切割后的人体实例图片大小转换成 224×224；同时本文将人体关节点坐标转换成大小为 56×56 的热度图，并将生成的热度图作为真实标签参与损失计算，实验损失函数为交叉熵误差函数，实验激活函数为 ReLU 函数，批处理大小为 128，学习率为 0.0006。CNPose 分为 3 个阶段，每个阶段深度分别为 2，2，4，每个阶段的多头自注意力过程的头部数目分别为 4，8，16，每个阶段 Patch Token 的维度分别为 128，256，512。

4.4 对比实验

本文选取了近年经典且先进的方法与本文提出的 CNPose 进行对比，其中包含 TFPose，HRNet，Simple Baselines。从表 1 中可知，在基础网络都选用 Resnet-50 时，CNPose 比 Simple Baselines 和 TFPose 在精度上分别高出 0.044，0.042；当基础网络选用 HRNet-W48 时，CNPose 比 HRNet 在精度上高出 0.016；当基础网络都是 Resnet 系列时，CNPose 采用 ResNet-50 比 Simple Baselines 采用 ResNet-101 和 ResNet-152 在精度上分别高出 0.03，0.023。从表 2 中可知，当基础网络采用 Resnet-50 时，CNPose 比 TFPose 在精度上高出 0.03；当基础网络采用 HRNet-W48 时，CNPose 比 HRNet 在精度上高出 0.002；当基础网络都使用 Resnet 系列时，CNPose 采用 ResNet-50 比 Simple Baselines 采用 ResNet-152 在精度上高出 0.015。综上所述，本文提出的耦合局部特征与全局表征方法相比其他方法效果更好。

表 1 在 COCO Val2017 数据集上的结果

Table 1 Experimental results on COCO Val2017 dataset

方法	基础网络	输入大小	AP	AP^{50}	AP^{75}	AP^m	AP^l
Simple Baselines	ResNet-50	384×288	0.722	0.893	0.789	0.681	0.797
Simple Baselines	ResNet-101	384×288	0.736	0.896	0.803	0.699	0.811
Simple Baselines	ResNet-152	384×288	0.743	0.896	0.811	0.705	0.816
HRNet	HRNet-W48	384×288	0.763	0.908	0.829	0.723	0.834
TFPose(Nd=6)	ResNet-50	384×288	0.724	—	—	—	—
CNPose	Resnet-50	224×224	0.766	0.935	0.837	0.733	0.813
CNPose	HRNet-W48	224×224	0.779	0.936	0.848	0.748	0.824

表 2 在 COCO dev-test2017 数据集上的结果

Table 2 Experimental results on COCO dev-test2017 dataset

方法	基础网络	输入大小	AP	AP^{50}	AP^{75}	AP^m	AP^l
Simple Baselines	ResNet-152	384×288	0.737	0.919	0.811	0.703	0.800
HRNet	HRNet-W48	384×288	0.755	0.925	0.833	0.719	0.815
TFPose(Nd=6)	ResNet-50	384×288	0.722	0.909	0.801	0.691	0.788
CNPose	Resnet-50	224×224	0.752	0.923	0.830	0.719	0.816
CNPose	HRNet-W48	224×224	0.757	0.925	0.833	0.720	0.817

4.5 消融实验

为了验证局部-全局信息交流模块对特征耦合的影响，本文在COCO2017 val 数据集上做了消融实验，对比实验组去掉了 CNPose 中的局部-全局信息交流模块，结果如表 3 所列。

表 3 消融实验结果

Table 3 Results of ablation experiment

方法	基础网络	AP	AP^{50}	AP^{75}	AP^m	AP^l	AR
CNPose	Resnet-50	0.766	0.935	0.837	0.733	0.813	0.793
CNPose-No-LGC	Resnet-50	0.762	0.935	0.827	0.728	0.809	0.788
CNPose	HRNet-W48	0.779	0.936	0.848	0.748	0.824	0.804
CNPose-No-LGC	HRNet-W48	0.776	0.936	0.848	0.745	0.823	0.802

在表 3 中，CNPose-No-LGC 代表去掉局部-全局信息交流模块的结构，模型输入大小为 224×224。从表 3 中可知，如果 CNPose 使用局部-全局信息交流模块，当基础网络分别是 Resnet-50 和 HRNet-W48 时，AP 指标分别能上涨 0.4%和 0.3%，实验证明了通过局部-全局信息交流模块来减少局部特征和全局表征的数据源范围差异有助于模型精度的提升。

结束语 针对前人结合局部特征和全局表征实现人体姿态估计的程度比较低的问题，本文提出了一个耦合局部特征和全局表征的串行网络 CNPose，该网络的局部-全局特征耦合模块利用注意力计算和残差的方式耦合卷积神经网络提取的局部特征和 Transformer 捕捉的全局表征。为了解决局部特征和全局表征在耦合过程中数据源范围存在差异带来的

模型精度丢失问题，本文提出了局部-全局信息交流模块，下采样各个块的特征图，将得到的特征图洗牌后加入到原特征图中，这种方式减小了两种特征图的数据源的范围差，从而提升了模型的精度。实验表明 CNPose 取得了比经典且先进模型更高的精度。但是本文提出的方法的参数量和计算量有待减少，这是作者未来需要研究改进的问题。

参 考 文 献

[1] IQBAL U，MILAN A，GALL J. Posetrack：Joint multi-person pose estimation and tracking[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu：IEEE，2017：2011-2020.

[2] HUANG S，GONG M，TAO D. A coarse-fine network for keypoint localization[C]// Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos：IEEE Computer Society，2017：3028-3037.

[3] PISHCHULIN L，INSAFUTDINOV E，TANG S，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：IEEE，2016：4929-4937.

[4] CAO Z，SIMON T，WEI S E，et al. Realtime multi-person 2d pose estimation using part affinity fields[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu：IEEE，2017：7291-7299.

[5] KRIZHEVSKY A，SUTSKEVER I，HINTON G E. Imagenet classification with deep convolutional neural networks[J]. Communications of the ACM，2017，60(6)：84-90.

[6] HE K，ZHANG X，REN S，et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas：IEEE，2016：770-778.

[7] SIMONYAN K，ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv：1409. 1556，2014.

[8] HOWARD A G，ZHU M，CHEN B，et al. Mobilenets：Efficient convolutional neural networks for mobile vision applications[J]. arXiv：1704. 04861，2017.

[9] TOSHEV A，SZEGEDY C. Deeppose：Human pose estimation via deep neural networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus：IEEE，2014：1653-1660.

[10] WEI S E，RAMAKRISHNA V，KANADE T，et al. Convolutional pose machines[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas：IEEE，2016：4724-4732.

[11] SUN K，XIAO B，LIU D，et al. Deep high-resolution representation learning for human pose estimation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach：IEEE，2019：5693-5703.

[12] MAO W，GE Y，SHEN C，et al. Tfpose：Direct human pose estimation with transformers[J]. arXiv：2103. 15320，2021.

[13] ZHANG Z，ZHANG H，ZHAO L，et al. Aggregating nested transformers[J]. arXiv：2105. 12723，2021.

[14] LIN T Y，MAIRE M，BELONGIE S，et al. Microsoft coco：Common objects in context[C]// European Conference on Computer Vision. Cham：Springer International Publishing，2014.

[15] RUGGERO RONCHI M，PERONA P. Benchmarking and error diagnosis in multi-instance pose estimation[C]// Proceedings of the IEEE International Conference on Computer Vision. Venice：IEEE，2017：369-378.

CHEN Qiaosong，born in 1978，Ph.D，associate professor. His main research interests include image processing，image understanding，artificial intelligence and computer vision.