引用格式：张大勇，陈一茗.尺度感知的多人姿态估计研究[J].中国传媒大学学报（自然科学版），2022，29(06)：50-57+67.
文章编号：1673-4793(2022)05-0050-09

尺度感知的多人姿态估计研究

张大勇*，陈一茗
（中国技术经济学会数字体育专业委员会，北京 100081）

摘要：近年，深度学习下的多人姿态估计研究取得长足进步，但如何应对场景中的尺度变化以及如何高效检测并分组多人姿态关键点仍是当前的巨大挑战。为提升网络对多人姿态的尺度感知能力，权衡模型的速度与精度，本文在关键点检测方面提出了一种尺度感知的多人姿态估计算法，结合高分辨率表征和变形感受野设计多人关键点特征提取模块，并更新迭代网络；同时在网络头部提出热力图指导的特征融合修正策略，丰富表征的多尺度表达。在Associate Embedding上应用自适应检测网络，MSCOCO数据集的定位精度提高了6.0%，表现出对困难姿势和中小尺度关键点的检测优势。

关键字：多人姿态估计；热力图；变形感受野；尺度感知
中图分类号：TP391　**文献标识码**：A

Research on scale-sensitiverepresentation for multi-person pose estimation

ZHANG Dayong*, CHEN Yiming
(Digital Sports Professional Committee,Chinese Society of Technology Economics, Beijing100081,China)

Abstract: In recent years, the research of multi-person pose estimation based on deep learning has made great progress. However, there remain huge challenges in coping with scale variation as well as efficiently detecting and grouping multi-person pose keypoints. In order to improve the scale sensibility of network and make a trade-off between speed and accuracy, we propose a scale-sensitive multi-person pose estimation algorithm involving keypoint detection network. Combining high-resolution representation and deformable receptive field, we design a multi-person keypoint feature extraction module and updated the network iteratively. Moreover, we propose a heatmap-guided feature fusion and refinement strategy to enrich multi-scale expression for the head of network. Applying adaptive detection network to the classic method Associate Embedding, with 6.0% localization improvement on MSCOCO validation accuracy, shows advantage on difficult poses and small scale keypoints detection.

Keywords: multi-person pose estimation; heatmap; deformable receptive field; scale-sensitive

1　引言

关键点检测网络是多人姿态估计研究的一大核心工作，能否准确定位多人关键点直接关乎姿态结果的精度高低。卷积神经网络是最常用且最强大的图像特征提取网络。Zeiler等人[1]将每层卷积输出可视化发现，随着卷积堆叠和网络加深，卷积网络抽取的特征从高分辨率低层次的边缘轮廓、方向细节和几何形态演变到低分

基金项目：国家重点研发研究计划(2018034)
作者简介(*为通讯作者)：张大勇(1990-)，男，高工，副主任委员兼秘书长，主要从事数字视频和大数据技术研究。Email：zhang_dayong@mixtmt.com

辨率高层次的语义抽象信息。自深度卷积发展以来，许多工作通过特征提取网络结构设计和多尺度特征融合技术，成功地增强了人体姿态估计模型对精细关键点的检测能力和对人体尺度变化的感知能力。

人体姿态估计任务对位置敏感度很高，基于热力图预测的网络中，特征和热力图的分辨率将直接影响最终的定位结果。近年，深度神经网络的发展启发了研究人员对网络结构的革新，许多工作开始从分辨率的角度思考如何提升关键点的表征能力。其中，HRNet[2]及其团队另一力作HigherHRNet[3]成功登顶当年自顶向下和自底向上多人姿态估计的榜单，并持续为后续工作提供方向与灵感。

本文先从分辨率和感受野两方面对现有的多人姿态关键点检测网络结构进行影响分析，然后提出多人关键点特征提取模块设计理念和特征融合修正策略。通过优化关键点检测网络，提升自底向上多人姿态估计方法的尺度感知性。

2 多人姿态关键点检测网络分析

以往关键点检测网络方法大多通过重复级联独立网络，多阶段地预测并修正同一学习目标。随着多人姿态估计问题研究的深入，关键点检测网络结构不断更新迭代。图1中给出了多人姿态关键点检测网络组成。首先，原始图像会在Stem部分进行分辨率调整，得到缩放后的网络输入。其次，通过特征提取网络进行图像语义的学习，获得关键点特征图。最后，特征图传入热力图预测网络估计关键点的位置，获得高斯响应热力图。整个关键点检测过程中，特征图和热力图在网络里前向传递、反向修正，不断训练。近年许多围绕特征图、热力图的改进工作卓有成效，下面分别从分辨率和感受野两方面对关键点检测网络进行影响分析。

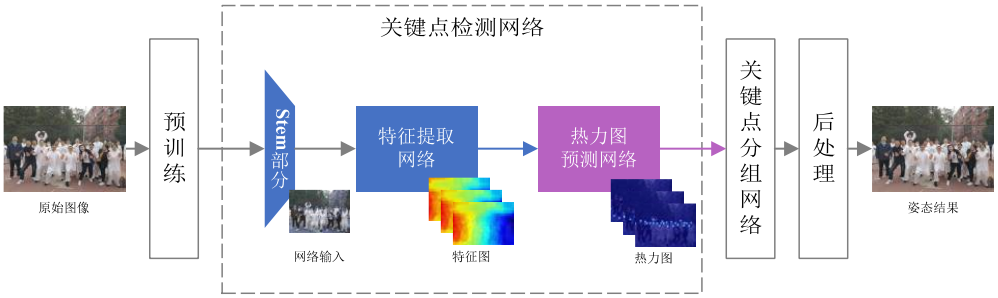


图1 多人姿态关键点检测网络组成部分示意图

2.1 分辨率对定位精度的影响

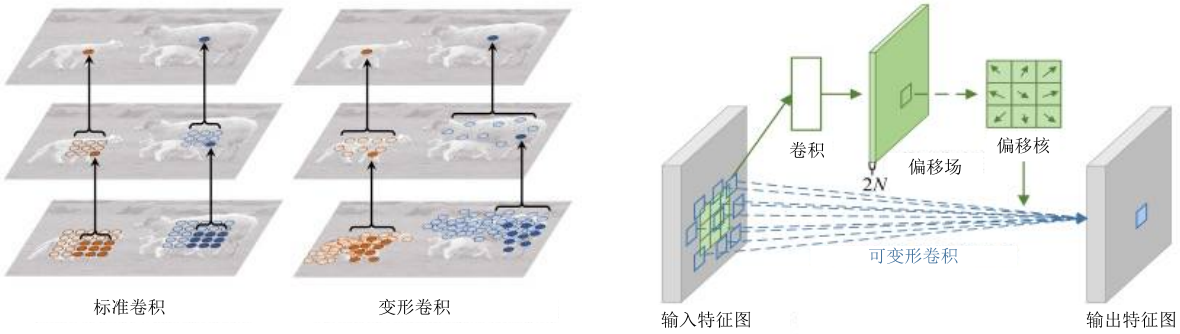
导致严重定位精度误差的原因之一是网络中的低分辨率表征。在特征金字塔理论里，高分辨率的表征一般可以保留更多的空间位置信息，而低分辨率的表征则能展现出更强的语义分析能力。因此许多经典工作一般从如何恢复高分辨率、如何维持高分辨率和如何融合多分辨率三方面入手。

PersonLab[7]方法简单粗暴，直接在输入网络前数倍放大原图，提升人体姿态估计模型的整体分辨率。同年，Magnify-Net[8]和Simple Baseline[9]通过对网络中的特征图进行线性插值或反卷积等上采样操作恢复高分辨率。然而，Sun [2]认为仅凭上述的上采样操作无法真正恢复有效的高分辨率特征，应该在不同语义阶段自始至终都维持高分辨率表征；同时受到Hourglass[4]和CPN[10]的多尺度连接思想启发，提出特征多次重复融合的高分辨率网络——HRNet[2]。其团队的另一力作HigherHRNet[3]则是同时针对特征图和热力图，利用HRNet提取高分辨率特征，再使用反卷积放大热力图的分辨率，最后提出多尺度热力图平均融合策略，大大增强了网络对尺度变化的鲁棒性。对于多尺度特征的融合方式除了上述提及的平均融合外，Su等人[11]也尝试在热力图上进行加权融合。关键点检测网络中的表征（特征图、热力图）分辨率对定位精度至关重要，因此选用高分辨率的网络设计往往可以事半功倍。

2.2 感受野对特征提取的影响

在卷积神经网络中，感受野（Receptive Field）指的是中层特征图上某神经元位置计算输出所用到的有效图像区域，示意图如图2(a)左图所示。相关工作[12, 13]认为，如果仅从局部细节出发，容易出现人体姿态的漏检和交叉误判；足够大的感受野可以包含更多的上下文信息，协助推理复杂场景下的多人姿态结果。

常见的增强感受野方式有使用更大的卷积核和采用扩张的空洞卷积(Dilated Convolution)[14],例如经典的算法CPM[15]采用9 × 9卷积配合多阶段级联网络来增大感受野,获得明显效果。类似地,循环(Recurrent)姿态网络[16]设计了一种循环递归模块来提升感受野。



(a)标准卷积和变形卷积感受野对比 (b)变形卷积示意图

图2 可变形卷积DCN示意图

由于人体姿态关键点尺度不一,处于较为精细位置的关键点需要较小的感受野才能捕获细节信息。因此一味地增大感受野不一定持续受益,反而会引入许多干扰信息。针对感受野的研究大体分为两类:第一,特征尺度金字塔;第二,几何变换自适应。可变形卷积DCN[17]的思想和实现过程如图2(b)所示,通过一个3 × 3卷积,对感受野上的每个卷积采样点学习相应的偏移量,使得常规的 $N \times N$ 卷积区域变形为不规则感受野,从而更好地拟合尺度不一的困难目标,与常规卷积的效果对比见图2(a)右图。然而,近年受其影响的多人姿态估计工作更倾向于在分组网络上迁移"偏移修正"概念,例如CenterNet[18]提出无锚偏移思想修正人体关键点;DEKR[19]参照空间变换网络STN[20]来设计自适应卷积(Adaptive Convolution)并构建多分支的关键点回归网络。

3 多尺度的自适应检测网络优化

有效的高分辨率表征和适度变形的卷积感受野对尺度不一的多人关键点检测大有裨益。结合多尺度的高分辨率网络和变形感受野思想,本节设计基于变形卷积的关键点检测模块DB-Module,并用优化后的模块批量更新高分辨率网络,配合热力图指导的特征融合修正策略,完成多尺度的自适应检测网络优化。

3.1 多人关键点特征提取模块设计

卷积网络发展至今,依靠更大、更多卷积的笨重设计已经暴露出明显缺点:计算量大且面临性能退化。Simonyan等人[21]提出使用多个3 × 3卷积代替较大卷积核,堆叠而成的感受野等大,同时引入更多非线性变换增强学习能力。He等人[22]针对深度网络的性能退化问题,推出跳跃连接的残差网络(Residual Network, ResNet)结构设计,利用残差学习思想缓解梯度爆炸和梯度弥散问题。

模块化设计这种"即插即用"的特征,使网络的改进变得简单快速。下面基于高分辨率网络HRNet[2]的主体部分,对每个阶段的子模块进行重新设计,并封装成DB-Module模块,然后批量替换整个网络,简单、快速地实现关键点网络的优化。

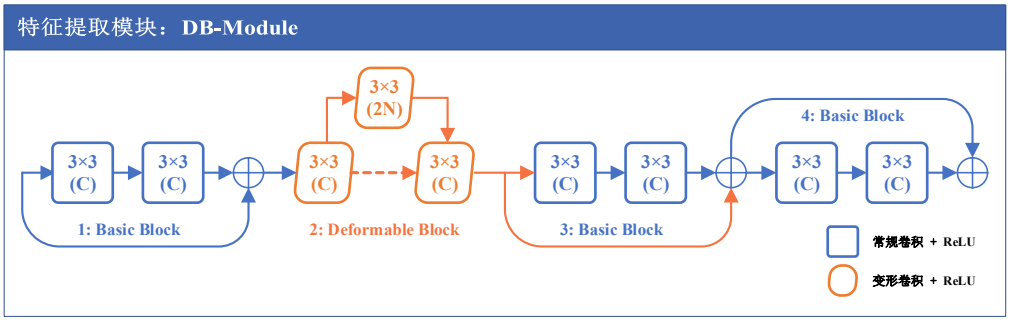


图3 特征提取模块DB-Module示意图

图3中展示了特征提取模块DB-Module的组成结构。DB-Module是本文关键点检测网络的基本模块，其中包含4个特征提取单元。模块中的特征提取单元分2种，蓝色方框部分采用残差结构[22]的Basic Block基础块；橙色变形方框部分则是以变形卷积[17]为灵感设计的Deformable Block变形块。卷积层的卷积核(Kernel))大小和通道数(Channel)分别表示为“$k \times k$”和“(C)”，空心块均由普通卷积/变形卷积和整流线性单位(Rectified Linear Unit, ReLU)[67]共同构成。

考虑到多人姿态场景中的复杂姿势和人体关键点的尺度变化，既需要足够大的感受野来适应变化的困难姿势，还应该保留较小的局部卷积区域来感知精细关键点。因此在DB-Module的模块设计中，仅允许1/4的特征提取单元进行不规则的感受野变形，同时采用跳跃连接减缓堆叠卷积造成的感受野发散和网络退化问题。参照经典目标检测工作[23, 24]，选取第2个连接单元进行变形操作可以将变形空间限制在整个模块的感受野中，让封装好的特征提取模块既保留高效的图像语义学习能力，又能发挥可变形卷积的尺度特性，更精准地捕捉困难人体实例。

3.2 热力图指导特征融合修正策略

优秀的多尺度表征不仅可以通过高分辨率网络和变形感受野提取得到，还可以利用特征融合策略进一步放大其尺度感知特性。Cheng等人[3]认为不同分辨率大小的高斯分布热力图可以“响应”不同尺度的人体关键点，因此在2020年提出了更高的高分辨率网络——HigherHRNet。其核心在于对热力图进行尺度增强，并在训练、推理阶段都使用多尺度融合策略，成功提升了中小尺度目标的解码定位精度。近期，该团队推出最新研究DEKR[19]，将高斯响应热力图的注意力机制特性与特征融合策略结合，通过热力图进行局部指导，也在定位精度上取得进步。

本节基于骨干网络HRNet[2]和分组方法AE[25]结合的多人姿态估计流程，提出一个简单的尾部融合策略：将热力图与高分辨率特征对齐平均相加后，再利用反卷积模块预测更高分辨率的热力图并在分组前对热力图进行融合修正(Aggregation Refine, A-Refine)。

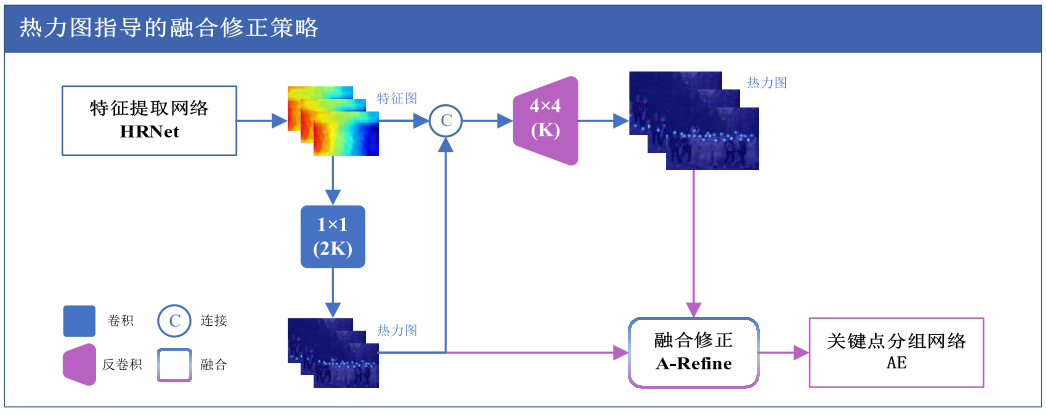


图4 热力图指导特征融合策略流程示意图

图4中可视化了热力图指导特征融合策略主要流程。由上一节的特征提取网络得到尺度感知特征图后，先照惯例通过一个1 × 1卷积层预测所需的热力图。一般该预测模块还同时预测标签集合用作分组关联信息指导，但分组算法不是本文重点，此处只形式化表示。按照尺度金字塔理论，低分辨率的热力图里含有较强的分类指导作用，再加上高斯响应本身自带的注意力机制，两者共同作用在尺度感知的特征图上可以融合成更强大的高分辨率表征，从而更精准地指导热力图预测。受到Simple Baseline[9]的启发，反卷积通常也被叫做转置卷积，通过反向捕捉卷积规律，既能够恢复部分有效的高分辨率表征，又可以在一定程度上拥有卷积的语义学习特性。本文沿用HigherHRNet[3]中反卷积层的结构设计与参数设定，在上采样出更大分辨率热力图的同时进行关键点的预测，并为多分辨率热力图设计A-Refine融合修正模块。以往工作中多使用连续的残差基础结构(Basic Res-Block)进行修正，本文额外增设变形模块DB-Module与DEKR[19]方法中的自适应矩阵(Adaptive Metrix)对比，从网络自行学习和手工主动设计两种改进角度寻求良好的修正模式。

本质上说，DCN[17, 24]和STN[20]均研究如何拟合物

体的空间几何变换,前者使用非参数式的网络自主学习思路进行模块级别设计,后者通过参数式的网络人工设计进行网络级别搭建。DCN方法易于泛化,即插即用但不可避免增加一定参数量;STN结构通过手动规划且在后续工作[19, 26]中被提炼成自适应矩阵(Adaptive Metrix)用于卷积改造,详见式(1)至(3),参数量可观但针对性强、不易泛化。

$$y(\boldsymbol{c}) = \sum_{i=1}^{9} \boldsymbol{w}_i x(\boldsymbol{o}_i + \boldsymbol{c}) \tag{1}$$

其中,$\boldsymbol{c} = (x_c, y_c)$ 表示中心(center)坐标,$\boldsymbol{w}_i$ 为卷积核的权值,$\boldsymbol{o}_i = (x_o, y_o)$ 表示距离中心的偏移量(offset)。其中 $\boldsymbol{o}_i$ 属于表示感受野偏移的2 × 9矩阵 $\boldsymbol{O}_t = \{\boldsymbol{o}_1, \cdots, \boldsymbol{o}_9\}$ 中元素。

DEKR将STN设计的矩阵放入MSCOCO[27]训练集中学习,获得整体的仿射变换矩阵 $\boldsymbol{A} \in \mathbb{R}^{2\times 2}$ 和翻转向量 $\boldsymbol{t} \in \mathbb{R}^{2\times 1}$。然后对常规(regular)卷积进行几何变换捕捉,求得变换(transformation)后的 $\boldsymbol{O}_t$,以下以3 × 3卷积为例:

$$\boldsymbol{O}_t = \boldsymbol{A}G_r + [\boldsymbol{t}\,\boldsymbol{t}\cdots\boldsymbol{t}] \tag{2}$$

$$G_t = \begin{bmatrix} -1 & 0 & 1 & -1 & 0 & 1 & -1 & 0 & 1 \\ -1 & -1 & -1 & 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix} \tag{3}$$

本文将可变形卷积和自适应矩阵应用到基于热力图预测的多人姿态估计中,通过在A-Refine融合修正模块上的实验对比择优,寻找良好的特征融合策略。

4 实验与结果分析

4.1 数据集及其评价指标

本文提出的关键点检测网络模块DB-Module和特征融合修正策略A-Refine均在MSCOCO[26]数据集上进行训练和验证。表1中给出了姿态估计任务常用的评价指标。

表1 人体姿态估计常用评价指标

评价指标	释义	数据集	描述
		单人姿态估计	
PCP	正确肢体的百分比	LSP	某阈值下,被正确检测的肢体部分比例;
PCK	正确关键点百分比	FLIC	某阈值下,被正确检测的关键点的比例;
		多人姿态估计	
OKS	目标关键点相似度	MSCOCO	类似目标检测里 *IoU*;
AP	平均精确度		•*mAP* 在 *OKS* = .50:.05:.95 时(主要指标) •AP^{50} 在 *OKS* = .50 时(宽松指标) •AP^{75} 在 *OKS* = .75 时(严格指标) •AP^{M} 为中等目标 $32^2 < area < 96^2$ 时 •AP^{L} 为大型目标 $area > 96^2$ 时
AR	平均召回率		•*mAR* 在 *OKS* = .50:.05:.95 时(主要指标) •AR^{50} 在 *OKS* = .50 时(宽松指标) •AR^{75} 在 *OKS* = .75 时(严格指标) •AR^{M} 为中等目标 $32^2 < area < 96^2$ 时 •AR^{L} 为大型目标 $area > 96^2$ 时

MSCOCO关键点挑战为人体姿态估计任务设计了一套多标准评价指标,以目标关键点相似度(Object Keypoint Similarity, OKS)系数和目标尺度为基准,计算平均精确度(Average Precision, AP)和平均召回率(Average Recall, AR)。

OKS 主要计算预测的姿态关键点与标注之间的相似度,数值在0~1之间。公式如下:

$$ks(\hat{\boldsymbol{p}}_i, \boldsymbol{p}_i) = e^{-\frac{\|\hat{\boldsymbol{p}}_i - \boldsymbol{p}_i\|_2^2}{2s^2k_i^2}} \tag{4}$$

$$OKS(\hat{\boldsymbol{p}}_i, \boldsymbol{p}_i) = \frac{\sum_i ks(\hat{\boldsymbol{p}}_i, \boldsymbol{p}_i)\delta(v_i > 0)}{\sum_i \delta(v_i > 0)} \tag{5}$$

此处,ks 为预测的人体的第 i 种姿态关键点坐标 $\hat{\boldsymbol{p}}_i$ 和实际的关键点坐标 $\boldsymbol{p}_i$ 的相似度;s^2 是当前人体分割掩码(Segmentation Mask)区域面积;$k_i = 2\sigma_i$ 为当前关键点的标注抖动分布归一化,用于调节关键点相对当前人体尺度的标注抖动。

AP 和 *AR* 针对预测中得分前20的姿态估计结果进

行计算，计算 OKS = .50:.05:.95 区域的 AP 和 AR 值，在不同的阈值下分别对两者求平均，可以得到最后使用的主流指标 $meanAP(mAP)$ 和 $meanAR(mAR)$。MSCOCO评价指标里还提供与尺度相关的指标 AP^M、AP^L 和 AR^M、AR^L。

本文提出的关键点检测模块设计和融合修正策略更关注于中小尺度人体关键点定位情况和召回能力的提升，而不是大尺度目标的评测情况，因此在后续的评估中，将针对与人物尺度相关的精细化评价指标 mAP、AP^M、mAR、AR^M 进行重点观测与分析。

4.2 实验设置及实施细节

本文使用Python语言和PyTorch深度学习框架实现基于变形卷积的关键点检测模块DB-Module，并批量更新自底向上的多人姿态模型HigherHRNet中关键点检测网络主体部分的HRNet，最后在分组网络AE[25]前置部分实现特征融合修正策略。

4.2.1 数据处理

在多人姿态估计网络的输入阶段，为便于数据并行，先对原图集体进行填充并缩放到512 × 512的固定尺寸，然后在训练阶段使用了随机裁剪、随机缩放、随机旋转和随机翻转等数据增广技术进行数据预处理。

根据本文3.2节提出的热力图指导特征融合修正策略，生成两种尺寸的Ground-Truth关键点热力图作训练标签，分别是128 × 128和256 × 256。

4.2.2 训练参数

随机初始化网络的权值，使用初始学习率为1.875e-3的Adam优化器对网络损失进行优化，训练的批尺度大小为8。学习率调整策略为先线性预热，后阶梯下降。本文实验训练了120个周期，总耗时约5.5天。学习率从 lr × 0.01 开始预热500个轮次，60个周期后开始阶梯式下降，在第80个周期降至1.875e-4。多次实验保留最佳验证结果并取平均。

整个网络在2张NVIDIA GTX 1080Ti GPU上进行分布式训练，同时采用线性尺度规则（Linear Scaling Rule）对基准网络预设的学习率进行调整，使之在不同批尺度大小和不同GPU数量的情况下，依旧获得接近原始训练精度的复现结果。

本文提出的特征融合策略中应用了两款融合修正模块：DB-Module和Adaptive-Metrix。两者均在模块更新后的网络上进行实验，各自分配1张NVIDIA GTX 1080Ti GPU并行训练。公平起见，模型微调（fine-tune）期间其余参数保持一致。整个实验过程中，前80个周期训练就变形模块DB-Module的DB-Pose网络，后40个周期内学习率 × 0.1并列进行最优融合修正模块的探索。

4.2.3 验证细节

本文的基准网络选定为HigherHRNet[3]，其分组算法沿用AE[25]。但是，由于HigherHRNet中自带多尺度热力图融合机制，容易混淆特征融合修正策略的有效来源，公平起见，本文将HigherHRNet中去除热力图融合策略后的主体网络HRNet[2]与分组算法AE拼合，作为第二基准网络进行参考。上述工作的源代码在验证、测试阶段均使用了 $[\times 0.5,\ \times 1,\ \times 1.5,\ \times 2]$ 尺度金字塔技术对预测的不同尺度的人体实例进行融合增强。为公平地验证尺度感知的关键点检测模块有效性，本文去除基准网络中的多尺度部分，并在本地环境下按与本文实验的相同配置重新运行和验证其开源模型，从而排除原文中额外进行姿态修正后带来的涨幅偏差。

4.3 实验结果及误差分析

本文以2020年榜首HigherHRNet作为第一基准网络；同时将2020年的多任务骨干网络HRNet和经典分组算法AE拼合，作为第二基准网络共同进行对比实验。为节省成本，仅在使用最小模型（w32）在MSCOCO[26]验证集上进行实验。

4.3.1 定量分析

在MSCOCO[26]验证集上的各项精细指标评测结果参见表2。验证集上的本地消融实验数据额外保留两位小数。

为便于区分，本文3.1节的DB-Module模块化设计对应模型表示为“DBPose”；后续加入本文3.2节特征融合修正策略A-Refine后，对应模型表示为“SSR-Pose”，进行如表2所示的消融实验。

(1) 关键点检测模块DB-Module的评测结果

根据表2中结果可知，单纯对基准网络HigherHRNet进行DB-Block批量更新得到的DBPose，无需微调即可得到66.83%的 mAP 精度，比第一基准模型提升1.14%。并且在各项与尺度变化相关的精细指标上都超越了基准，其中 AP^M 和 AR^M 较为明显，分别是1.41%和1.45%的涨幅，体现了DB-Module变形感受野在捕捉尺度不一人体方面的优势。

表2 MSCOCO验证集上的结果

方法	骨干网络	AP(%)					AR(%)				
		mAP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	*mAR*	AR^{50}	AR^{75}	AR^{M}	AR^{L}
HRNet + AE	HRNet-w32	61.88	85.20	67.45	55.85	71.37	67.99	88.38	72.69	60.23	78.75
HigherHRNet	HRNet-w32	65.69	85.85	71.33	59.99	74.27	70.56	88.11	75.33	63.54	80.44
DBPose	SSR-w32	**66.83**	86.12	72.75	**61.40**	75.76	**71.69**	88.71	76.62	**64.99**	81.28
CHED	HRNet-w32	66.97	86.17	72.70	61.43	75.51	71.65	88.79	76.28	64.99	81.12
SSR-Pose	SSR-w32	**67.83**	86.84	73.31	**62.35**	76.12	**72.57**	89.29	77.06	**66.11**	81.66

采取特征融合修正策略A-Refine将两者融合形成本文的尺度感知多人姿态估计模型SSR-Pose，通过下述消融实验探索“性价比”更高的最终模型。表2中的最后实验条目，整体平均精度 *mAP* 达到67.83%。与第一和第二基准网络相比，分别提升2.14%和5.95%的平均精度，尺度指标 AP^{M} 上的涨幅更是高达2.36%和6.50%。

(2) 特征融合修正模块A-Refine的消融实验

表3中分别对三种类型的特征融合修正模块进行实验评测，实验条目1，2为基于STN和DCN设计的Adaptive Metrix和DB-Module。实验条目3为使用HR-Module的基础模式搭建的高分辨率修正模块。

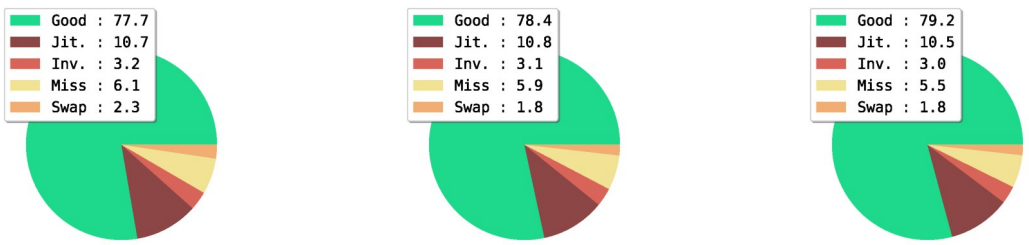
表3 特征融合修正模块的消融实验

序号	Adaptive Metrix	DB-Module	HR-Module (Basic)	AP(%)					AR(%)				
				mAP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	*mAR*	AR^{50}	AR^{75}	AR^{M}	AR^{L}
1	√			67.34	86.66	73.05	61.94	75.76	72.42	89.17	77.08	66.09	81.46
2		√		67.83	86.85	73.46	62.31	**76.31**	72.54	89.11	77.08	65.87	**81.98**
3			√	**67.83**	86.84	73.31	**62.35**	76.12	**72.57**	89.29	77.06	**66.11**	81.66

从实验数据上看，手工设计的Adaptive Metrix自适应矩阵在微调的情况下仍需要更复杂的参数调整才能获得理想精度，泛化性和拓展性不强。本文提出的DB-Module虽能得到与HR-Module高分辨率修正模块相同的精度，但对比尺度指标发现，变形感受野更擅长捕捉困难的大型人体，而高分辨率的基础模块才更适合用于修正精细关键点坐标的偏移；同时变形卷积DB-Module代码量稍大，耗时略久，因此性价比更高的方法为使用HR-Module (Basic)。本文SSR-Pose的最终版本搭建拟使用基于高分辨率的特征融合修正模块。

4.3.2 误差分析

采用coco-analyze误差分析工具[28]对基准网络HigherHRNet[3]和本文的变形感受野检测网络DBPose以及应用特征融合修正策略后的SSR-Pose进行定量评价误差分析。定位误差的结果是从被成功检测的姿态关键点中求得，与验证集评测结果有所出入，因此以下主要对比分析误差趋势。



(a)HigherHRNet (b)DBPose (c)SSR-Pose

图5 不同类型定位误差得分布情况

图5中三种方法在定位误差上的分布大致相同，均有较高的Jitter抖动误差，和较小的Swap交换误差和Inversion逆转误差，具体数值结果和比较见表4。

表4 四类定位误差的数值结果与趋势

序号	方法	骨干网络	Good (%)	Localization Errors (%)			
				Jitter	Inversion	Miss	Swap
1	HigherHRNet	HRNet-w32	77.7	10.7	3.2	6.1	2.3
2	DBPose	HRNet-w32	78.4↑	10.8↑	3.1↓	5.9↓	1.8↓
3	SSR-Pose	SSR-w32	**79.2**↑	**10.5**↓	**3.0**↓	**5.5**↓	**1.8**↓

通过条目1和2的对比发现，使用变形感受野模块DB-Module更新高分辨率的关键点检测网络，可以明显缓解Miss遗漏误差和Swap交换误差。这说明对本文提出的特征提取模块优化同时拥有高分辨率和尺度感知的特性，共同作用缓解因分辨率变化带来的定位丢失问题；同时得益于更高质量关键点预测热力图，其分组效果也有所提升。对比条目2和3，最终的SSR-Pose通过热力图指导的融合修正策略，对重点的局部精细区域投入更多注意力，使得最终预测的关键点热力图具备更强尺度感知能力，因而有更小的Miss遗漏误差和Jitter抖动误差。Good优秀分类指标得分大幅提升，最终SSR-Pose的每项定位误差指标均低于基准网络，较难察觉、看似影响较小的Jitter抖动误差都获得明显的缓解。可见，本文提出的自适应检测网络在困难姿势和精细关键点的检测上具有优势。

5 结论

本文通过分析高分辨率和变形感受野对网络性能的影响，设计一款基于可变形卷积的特征提取子模块。通过模块化的设计批量更新迭代网络架构，实现特征提取骨干网络的优化。为增强整体结构的尺度感知能力，在任务头部处提出了一个简单的尾部融合策略，利用网络中增强的高分辨率热力图指导特征，配合特征融合修正模块，共同完成尺度感知的关键点自适应检测网络优化，丰富了表征的多尺度表达，表现出对困难姿势和中小尺度关键点的检测优势。

参考文献（References）：

[1] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks[C]//European Conference on Computer Vision, 2014 :818-833.

[2] Sun K, Xiao B, Liu D, et al. Deep high-resolution representation learning for human pose estimation[C]// IEEE Conference on Computer Vision and Pattern Recognition, 2019.

[3] Cheng B, Xiao B, Wang J, et al. HigherHRNet: scale-aware representation learning for bottom-up human pose estimation[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.

[4] Newell A, Yang K, Deng J. Stacked hourglass networks for human pose estimation[C]//European Conference on Computer Vision, 2016 : 483–499.

[5] Li W, Wang Z, Yin B, et al. Rethinking on multi-stage networks for human pose estimation[DB/OL]. arXiv: 1901.00148, 2019.

[6] Wang J, Sun K, Cheng T, et al. Deep high-resolution representation learning for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020,43(10): 3349 - 3364.

[7] 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]//European Conference on Computer Vision (ECCV), 2018.

[8] Wang H, An W P, Wang X, et al. Magnify-net for multi-person 2d pose estimation[C]//IEEE International Conference on Multimedia and Expo (ICME), 2018 .

[9] Xiao B, Wu H, Wei Y. Simple baselines for human pose estimation and tracking[C]//European Conference on Computer Vision (ECCV), 2018: 472–487.

[10] Chen Y, Wang Z, Peng Y, et al. Cascaded pyramid network for multi-person pose estimation[C]//IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[11] Su Z, Ye M, Zhang G, et al. Cascade feature aggregation for human pose estimation[DB/OL]. arXiv: 1902.07837 , 2019.

[12] Luo W, Li Y, Urtasun R, et al. Understanding the effective receptive field in deep convolutional neural networks[C]//30th International Conference on Neural Information Processing Systems, 2016.

[13] Li J, Wang Z. Real-time traffic sign recognition based on efficient CNNs in the wild[J]. IEEE Transactions on Intelligent Transportation Systems, 2018, 20(3): 975-984.

[14] Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions[DB/OL]. arXiv:1511.07122 ,2015.

[15] Wei S-E, Ramakrishna V, Kanade T, et al. Convolutional pose machines[C]//IEEE conference on Computer Vision and Pattern Recognition, 2016.

[16] Belagiannis V, Zisserman A. Recurrent human pose estimation[C].//12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), 2017 .

（下转第67页）