DOI: 10.13382/j.jemi.B2306196

改进 KAPAO 的人体关键点检测*

赵 普[1] 武 一[1,2]

(1. 河北工业大学电子信息工程学院 天津 300401;2. 河北工业大学电子与通信工程国家级实验教学示范中心 天津 300401)

摘 要:针对人体关键点检测存在检测精确度低的不足,在 KAPAO(keypoints and pose as objects)网络的基础上进行改进。使用 PoseTrans(pose transformation)进行数据增强,提高网络的泛化性;针对特征融合能力的不足,设计融合注意力机制的 BiFPN(Bi-directional feature network)模块充分融合不同语义特征,提高网络对深层语义信息和浅层语义信息的融合能力;在网络输出阶段设计自适应扩张卷积模块,将不同扩张率的输出分支进行自适应融合,有效获得图像的全局信息;在网络的后处理部分设计 SDR-NMS(soft DIOU relocation non-maximum suppression)替代传统的 NMS,保留最优的关键点预测框。实验结果表明,网络的 AP 分数提高了 4.8%,AP 为 68.6%,检测速度为 19.1 ms。网络精确度和检测速度均具有较好的表现性。

关键词: 关键点检测;非极大值抑制;注意力特征融合;数据增强

中图分类号: TP391.41;TN99 **文献标识码**: A **国家标准学科分类代码**: 520.20

Improved human keypoints detection for KAPAO

Zhao Pu[1] Wu Yi[1,2]

(1. School of Electronics Information Engineering, Hebei University of Technology, Tianjin 300401, China; 2. National Demonstration Center for Experimental, Electronic & Communication Engineering, Hebei University of Technology, Tianjin 300401, China)

Abstract: For the lack of detection accuracy for human keypoints, it is improved on the basis network of KAPAO (keypoints and pose as objects). The generalization of network is improved by the enhance data method of PoseTrans (pose transformation); for the lack of characteristic fusion capabilities, the BiFPN (Bi-directional feature network) module is designed to fully integrate different semantic characteristic to improve the integration ability of deep semantics information and shallow semantic information; the adaptive expansion convolution module is designed to adaptive fusion different expansion rates of output branch during the network output phase, it effectively obtains the global information of the image; in order to retain the optimal key point prediction box, the traditional NMS is replaced by SDR-NMS (soft DIOU relocation non-maximum suppression) during the post-processing part of the network. The experimental results show that the AP score was increased by 4.8%, the AP was 68.6%, and the detection speed was 19.1 ms. The accuracy and detection speed of network have better performance.

Keywords: keypoints detection; non-maximum suppression; attention feature fusion; data augmentation

0 引 言

人体关键点检测是计算机视觉重点研究领域之一,其目的是从输入的图像中检测出鼻子、手腕、臀部、脚踝等人体关键点,获得各个人体关键点的位置坐标,从而有助于行为识别[1-2]、人机交互、体育运动分析等下游任务的实现。

随着深度学习的发展,研究人员将深度学习与人体关键点检测技术结合取得了相应的进展,但是同时存在推理速度慢、在具有挑战性的场景下,如罕见姿态以及多尺度输入时,检测精度低等问题。现阶段基于深度学习

收稿日期: 2023-01-13 Received Date: 2023-01-13

*基金项目:国家自然科学基金(51977059)、河北省自然科学基金(E2020202042)项目资助

的人体关键点检测可分为基于热图和基于关键点坐标回归两类方法。其中较为常用的方法是基于热图的方法。Cao 等[3]提出使用部分亲和场将身体部位与图像中个体相关联,通过全局遍历解析热图和亲和场的对应关系,得到所有人体关键点。但是算法在多人场景下会出现关键点与人体部位匹配错误的问题。Chen 等[4]针对遮挡的关键点提出一种级联金字塔的新型网络结构,使用 GlobalNet(global network)检测简单关键点,再通过 RefineNet(refined network)进一步细粒度检测图像中的关键点信息。Wang 等[5]提出一种高分辨率的网络称为 HRNet(hight-resolution network),在整个网络过程中保持高分表率表示,并将高分辨率与低分辨率的卷积通道进行并行连接,得到表示更精确的空间特征图。由于网络始终保持高分辨率,所以模型参数量和复杂度较大。为此,后续提出一系列基于 HRNet 的改进以解决模型参数量大的问题。Yu 等[6]提出 Lite-HRNet 网络,将 shuffle 模块[7]融入 HRNet,并使用信道加权取代高昂的逐点卷积。钟宝荣等[8]提出在 HRNet 网络中融入 Ghost 模块[9],Sandglass 模块以及注意力机制的轻量关键点检测网络。马皖宜等[10]提出融入多谱注意力机制的 Lite-HRNet,在网络后接入深度可分离的反卷积提升网络检测速度。

以上基于热图的方法存在量化误差,当两个相同关键点相邻较近时,可能会被误认为是一个关键点。同时,关键点检测精确度受到热图分辨率的限制,导致检测速度较低。因此,研究人员开始研究基于关键点坐标直接回归的方法,摒弃热图以提高检测速度。Toshev 等[11]提出基于关键点坐标回归的检测方法,将深度学习引入关键点检测替代基于模板匹配的方法。Zhou 等[12]将关键点对象作为检测边界框的中心点,使用 CenterNet 检测器进行关键点定位。Li 等[13]提出一种带有残差对数似然估计的新回归范式来捕捉潜在的输出分布,摒弃热图并与现有的关键点检测网络耦合,达到了一定速度的提升。Nie 等[14]提出一种 SPM(single-stage multi-person pose machine)网络,将关键点编码为相对于根关节的位移。通过网络同时预测根关节和关键点位移。但当出现人物严重遮挡时,其表现较弱。Li 等[15]为了解决基于回归的方法准确率较低的问题提出一种级联的 Transformer 关键点检测网络,多层的自注意力用于捕捉各个关键点之间的空间关系。

综上所诉,基于热图的方法虽然精度较高,但是其后处理需要大量时间,推理速度缓慢;基于关键点回归的方法摒弃热图回归以提高推理速度,但其检测精度较低,为此,针对基于关键点回归的检测网络精确度较低的问题,本文基于 KAPAO-S[16]网络做出改进,使用 PoseTrans[17]数据增强方法提高网络的泛化性,设计融入注意力融合机制的 BiFPN 模块,在特征融合部分融入设计的空间注意力机制,提高特征融合能力,并针对多尺度目标,融合不同扩张率的扩张卷积提高对多尺度目标的检测能力;最后使用改进的 NMS 算法(SDR-NMS)提高关键点的定位精确度。

1 相关工作

KAPAO 模型是一种多任务损失模型,通过模型同时检测一组关键点对象和一组人体姿态对象。首先将关键点检测转化为基于锚框的目标检测。关键点坐标为正方形边界框的中心,使用归一化的方法设置边界框大小。同时对各类关键点坐标和人体边界框进行空间建模,对各类关键点坐标值和人体边界框进行回归,进而估计出人体姿态对象。关键点对象用于检测局部特征较强的单个关键点,比如眼睛耳朵等。关键点检测的最终输出为关键点检测框,进而转化为关键点坐标值和其置信度。人体姿态检测用于检测人体姿态,最终输出为人物检测框和估计出的 17 个关键点坐标值。如图 1 所示,人体姿态对象 O^p,包含一个人体对象和 17 个关键点坐标,关键点置信度为 0。其输出框的大小为 57,分别包含人体边界框信息($\boldsymbol{p}_o$,$\boldsymbol{t}_x$,$\boldsymbol{t}_y$,$\boldsymbol{t}_w$,$\boldsymbol{t}_h$),1 个人体类别($\boldsymbol{c}_1$),17 个关键点的类别($\boldsymbol{c}_2$,…,$\boldsymbol{c}_{k+1}$),其中 k = 17,17 个关键点坐标($\boldsymbol{v}_{x1}$,$\boldsymbol{v}_{y1}$,…,$\boldsymbol{v}_{xk}$,$\boldsymbol{v}_{yk}$)。关键点对象 O^k,虽然其输出框的大小为 57,但是仅用关键点边界框信息($\boldsymbol{p}_o$,$\boldsymbol{t}_x$,$\boldsymbol{t}_y$,$\boldsymbol{t}_w$,$\boldsymbol{t}_h$)和 17 个关键点的类别($\boldsymbol{c}_2$,…,$\boldsymbol{c}_{k+1}$)作为关键点检测的信息进行损失计算。关键点对象作为单独的对象无需与其余关键点进行建模。算法使用共享的检测头同时检测这两种对象。并使用简单的匹配算法将单个关键点对象匹配到对应的人体部分,从而避免使用复杂的自底向上的分组方法。即当预测的关键点对象坐标值与预测的人体姿态对象中关键点坐标值相匹配且关键点置信度大于设定的阈值时,则将其匹配到检测的人体姿态对象中。算法以 YOLOv5 的 3 种算法模型为基础,共有 KAPAO-S,KAPAO-M,KAPAO-L 这 3 种模型。本文选用以 YOLOv5s 为基础的 KAPAO-S 模型,整体算法结构如图 2 所示,其中 O^p 为姿态对象,为 O^k 关键点对象,ϕ 为关键点与姿态匹配算法,NMS 为非极大值抑制算法。

2 本文方法

本文使用 PoseTrans 特征增强方法生成更多罕见的姿态,以扩充人体姿态的多样性,提高模型对于弱局部特征关键点检测的泛化性,对于 PANet[18]在特征提取时产生冗余的特征图,特征融合不充分的问题,设计一种注意力机制的 BiFPN[19],BiFPN 特征金字塔网络通过跳跃连

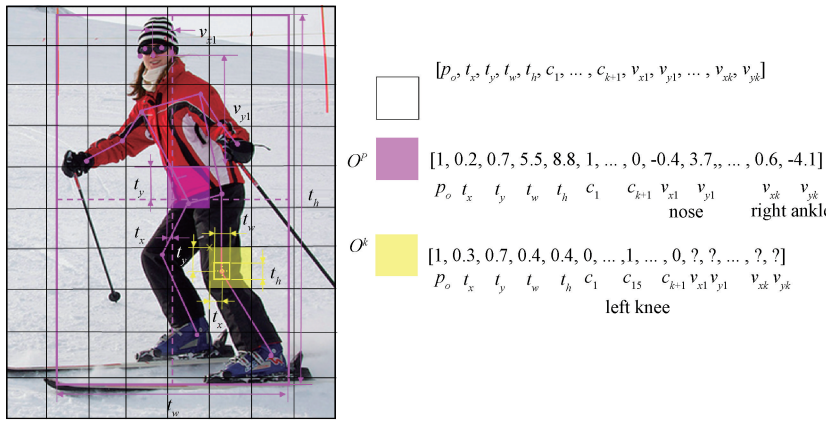

图 1　姿态对象和关键点对象示意图

Fig. 1　Schematic figure of pose and key object

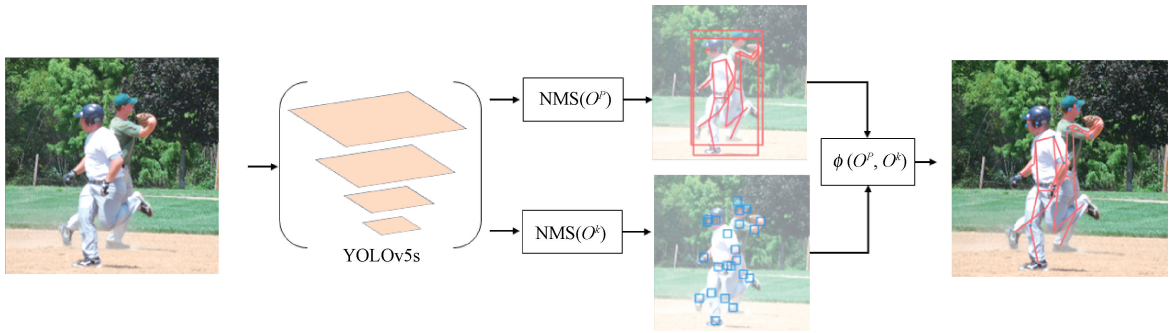

图 2　算法结构图

Fig. 2　Algorithm structure diagram

接和特征通道加权有效减少了冗余特征，并在 BiFPN 的基础上融入本文的注意力融合机制模块，称为 A-BiFPN。不同的特征图不再是简单的拼接，而是获取不同特征图的空间权重，增强特征融合能力。在 Neck 部分后加入自适应扩张卷积模块，分别对应 Neck 部分的 4 层输出。其由 4 个自适应扩张卷积模块 ADC（adaptive dilatedconvolution）组成，由不同的扩张率的扩张卷积与原始特征图并联，自适应的获得不同尺寸的感受野，提高模型对多尺度目标的检测能力。在后处理部分对于 NMS 算法将相邻的预测框置信度分数强制置零，导致漏检的问题，提出本文的 SDR-NMS 算法，分别经过置信度抑制和重定位，获取最优的预测框。改进后的网络结构如图 3 所示。

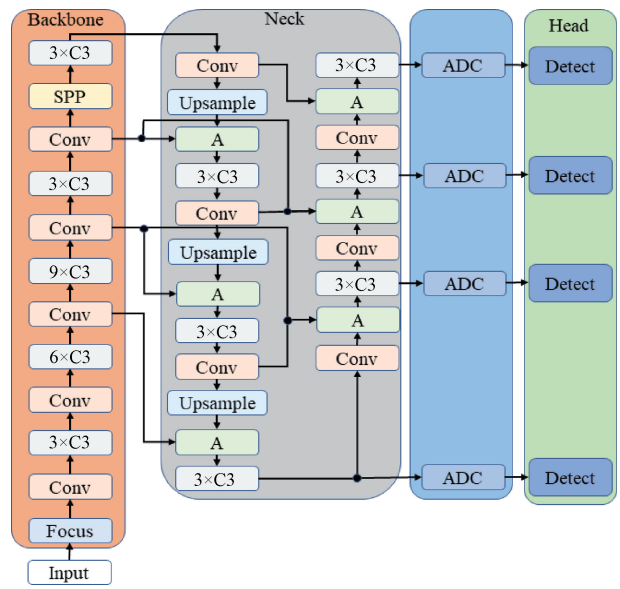

图 3　改进特征提取网络

Fig. 3　Improved feature extraction network

2.1　数据增强

通过对 COCO 关键点数据集进行归一化并聚类为 20 个姿态类别发现，数据集存在长尾分布现象，如图 4 所示，其中，站立，行走等常见的姿态占据了数据集的大部分，而罕见的蹲下等姿态仅占较小部分。

为此，使用 PoseTrans 数据增强方法对 COCO 数据集进行扩充。PoseTrans 数据增强的方法如图 5 所示，以 3 类姿态为例：首先将人体拆解为头部，躯干和四肢。使用

PTM(pose transformation module)姿势变换模块通过仿射变换对四肢进行旋转和缩放生成多样的姿势。再通过姿势鉴别器 D,去除关节角度扭曲,位置不合理的姿态,形成候选姿态池。图5中 P_1,P_2,P_3 分别为对应候选姿态池中的3种姿态的概率。最后通过 PCM(pose clustering module)聚类算法,对原始 COCO 数据集聚类得到每类对应的类别权重 α_L,L 表示类别。预测生成的每一个新的姿态属于每个类别的概率 P_n^L,n 表示生成的总的姿态个数($n \in \{1,2,3\}$),表示属于某一类别($L \in \{A,B,C\}$)。

最后通过选取加权和最小的姿态作为扩充的罕见姿态,用于对数据集的增强,以 P_2 为例,加权求和计算公式如式(1)所示,并通过式(2)求出概率最小的值为 P_2,即概率为 P_2 的姿态为最终的姿态。如图5中姿态池中边界框所标记的姿态。

$$P_2 = p_2^A\alpha_A + p_2^B\alpha_B + p_2^C\alpha_C \tag{1}$$

$$t = \min(P_1,P_2,P_3) \tag{2}$$

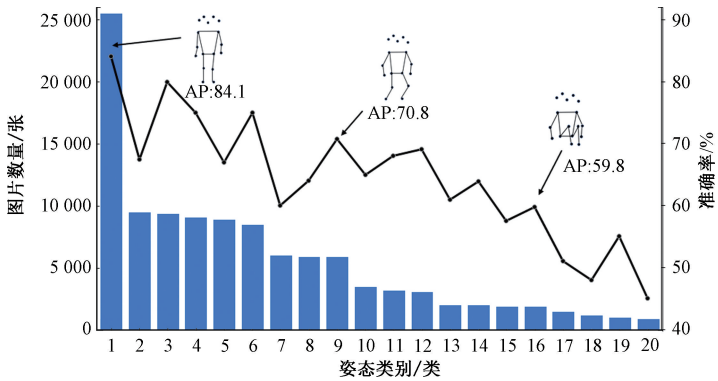

图4　姿态类别频率和平均准确度分布

Fig. 4　Pose categories frequency and average accuracy distribution

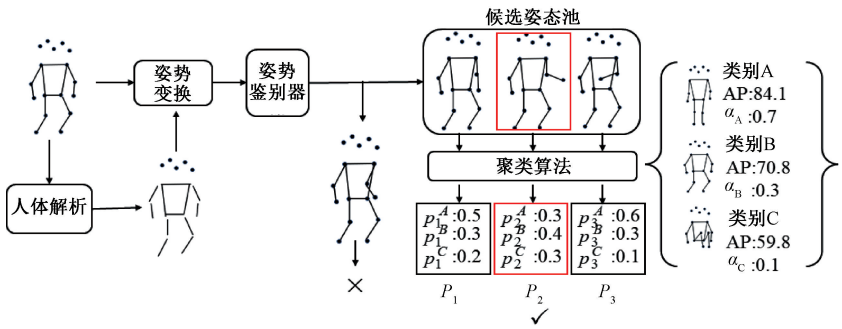

图5　PoseTrans 算法结构

Fig. 5　PoseTrans algorithm structure

2.2　注意力融合机制的 BiFPN

为了对主干网络提取的不同特征层充分的融合,首先提出 FPN[20] 融合结构,增加一条单向自底向上的路径,对不同尺度的特征图进行简单的上采样。PANet 在 FPN 的基础上引入一条自上而下的路径,由于反复的特征提取,产生了冗余的特征图降低了网络效率。BiFPN 将只有一个输入边的特征层进行残差连接,对特征融合结构进行简化,同时为了增强特征融合能力,将处于同一级的特征层,从原始输入节点增加一条跳跃连接,使得网络同时融合了深层和浅层的特征。其具体结构如图6(d)所示,由于 BiFPN 等特征融合网络在融合节点是进行简单的拼接,在融合节点设计注意力融合模块,增强特征融合的能力。

对不同特征图使用空间注意力机制,得到不同空间权重的特征图,新的特征映射与原始的特征图相结合。注意力机制融合部分结构如图7所示,以跳跃连接的3层特征为例,其中 a,b,c 分别是不同的特征图,分别进行最大池化和1×1卷积,1×1卷积用于保留全局信息,最大池化用于获取局部信息,并对两者进行拼接再通过3×3卷积,最后通过 Sigmoid 函数,为每个特征图生成相应的空间权重,生成的权重图和原始特征图进行矩阵运算,最后融合具有丰富的语义信息。

2.3　自适应扩张卷积模块

Neck 的后部加入不同扩张率的扩张卷积,根据图像的不同尺度自适应学习各特征图中不同的感受野,从而提高不同尺度关键点检测的精确度。如图8所示,首先

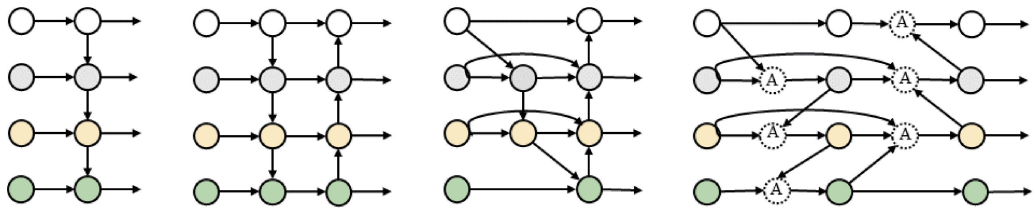

图 6　特征金字塔结构对比

Fig. 6　Comparison of feature fusion structures

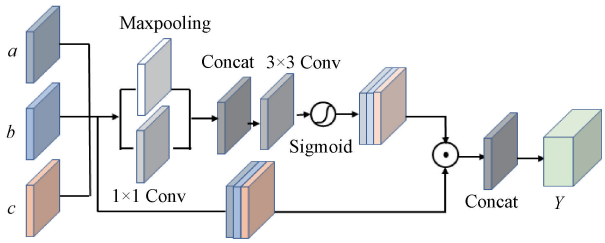

图 7　注意力融合机制结构

Fig. 7　Structure of attention fusion mechanism

将不同扩张率的分支与原始特征图分支并联,不同扩张率分支提供不同大小感受野的特征图,然后利用分支池化层自适应融合 3 个不同分支的特征图信息,提高多尺度预测精确度。3 支路分支由一个常规卷积分支和两个不同扩张率的扩张卷积的分支组成,每个分支的卷积操作后为 BN 层和 ReLU 激活层。3 个支路具有相同大小的 3×3 卷积核,其中扩张卷积的扩张率分别设置为 3,5。

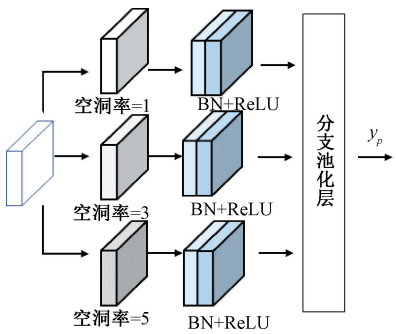

图 8　自适应扩张卷积

Fig. 8　Adaptive dilated convolution

分支池化层融合来自不同并行分支的信息,并避免引入额外的参数。平均运算用于在训练期间平衡不同并行分支的表示,表达式如式(3)所示。

$$y_p = \frac{1}{D_b}\sum_{i=1}^{D_b} y_i \tag{3}$$

其中, y_p 是分支池化层的输出, D_b 是分支数量,即 $D_b = 3$。

2.4　SDR-NMS 算法

在检测任务中,通常使用 NMS 算法,对预测框进行保留和抑制。传统的 NMS 算法置信度更新过程如式(4)所示:

$$s_i = \begin{cases} s_i, iou(M, b_i) < N_i \\ 0, iou(M, b_i) \geqslant N_i \end{cases} \tag{4}$$

式中: M 为经过置信度排序选出的最优预测框, b_i 为预测框,当两者的 iou 大于等于阈值 N_t 时,将其对应的置信度 s_i 置为 0, M 框将 b_i 框抑制并舍弃。

在人员拥挤的场景下,关键点对象之间将会存在相互遮挡的现象,不同关键点对象的预测框重叠度较高,传统的 NMS 将会发生漏检,定位精确度不准确,从而影响检测精确度。于是,本文设计一种 SDR-NMS 算法。该算法分为置信度抑制和预测框重定位两阶段。受 Soft-NMS[21] 启发,利用置信度抑制的方式调整类别的置信度并重新排序。选出置信度得分最高的作为最优预测框 M, 并记录 DIOU 小于 N_t 的预测框的位置信息,最后利用位置信息对最优预测框进行重定位。

在置信度抑制阶段,对最优预测框 M 重叠率较高预测框使用软抑制机制,适当的降低其置信度,得到新的置信度集合 Z,并根据集合 Z 对预测框集合重新排序得到新的置信度集合 B, 预测框置信度公式如式(5)所示。

$$s_i = \begin{cases} s_i, D_{iou}(M, b_i) < N \\ s_i \times (1 - D_{iou}(M, b_i)), D_{iou}(M, b_i) \geqslant N \end{cases} \tag{5}$$

式中:使用 D_{iou} 作为预测框的阈值判断准则。

在重定位阶段,首先找到与最优预测框的 DIOU 值小于阈值的 k 个预测框集合 $A(A\{a_1, a_2, \cdots, a_k\})$,然后计算最优预测框和集合 A 之间的平均偏移值,最后利用平均偏移值进行重定位操作,获得最终的最优预测框 J。重定位计算过程如图 9 所示。

计算最优预测框 M 的 x_m 值与集合 A 中 a_i 预测框对应的 x_a 值的偏移值 O_x ,并使用 x_m 与偏移值相加得到最终得预测框 J 的 x_i 值,以预测框的左上角坐标值 x_i 重定位为例计算公式如下所示,以此类推便可求得经过重定位计算之后的预测框 J。

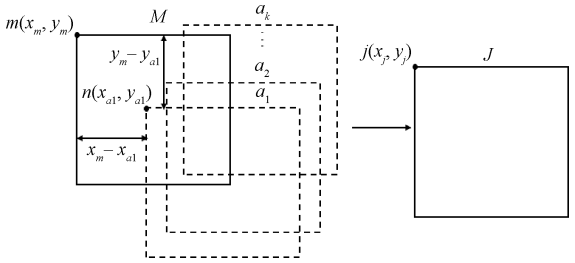

图 9　重定位计算过程

Fig. 9　Figure of relocation calculation process

$$O_x = \frac{\sum_{i=1}^{k} | x_m - x_{a_i} |}{k} \tag{6}$$

$$x_j = x_m + O_x \tag{7}$$

3　实验结果与分析

实验环境配置如下:操作系统为 Unbunt18.04,深度学习框架为 Pytorch1.9,CUDA 版本为 CUDA11.0,GPU 为 RTX3060(12 G)×4,编程语言为 Python3.7。模型训练过程中,输入数据大小为 640×640,batch-size 设为 96,epoch 为 1 000。使用 Nesterov 动量的随机梯度下降优化器,在学习率调整过程中,前 5 个 epoch 使用 warm-up 策略[22],后使用 SGDR 策略[23]。

3.1　数据集介绍

选用 COCO2017 数据集作为实验数据集进行训练验证和测试,该数据集 train2017 训练集含有 118 287 张人体姿态图片,val2017 验证集包含 5 000 张图片,test-dev 测试集包含 20 000 张图片,含有 17 个人体关键点。

3.2　边界框生成

本文使用关键点坐标作为边框的中心值为 17 类关键点生成边界框。生成 5 种不同尺寸的边界框 $b_s \in \{S\times1\%, S\times2.5\%, S\times5\%, S\times7.5\%, S\times10\%\}$ 经验证使用 S×2.5% 以下尺寸的边界框使得模型训练不稳定,准确性较差;当尺寸大于 S×5%时,准确性会迅速下降。使用 $S\times5\%$ 大小的边界框尺寸,模型性能表现最优,因此本文选择的边界框大小为 $S\times5\%$。

3.3　评价指标

对于关键点检测任务使用常用的关键点相似度评价指标,其公式如式(8)所示。

$$OKS = \frac{\sum_i \exp\left(-\frac{d_i^2}{2s^2k_i^2}\right)\delta(v_i > 0)}{\sum_i \delta(v_i > 0)} \tag{8}$$

其中,d_i 表示关键点预测值与真实值的欧氏距离,s 表示目标尺寸,k_i 表示为一个用于控制每类关键点衰减的常数,v_i 表示关键点的可见性标志。采用平均精确度 AP 和召回率 AR 作为检测的评价指标。当 OKS 值不同时,得到不同的 AP 和 AR,OKS 可取 {0.50,0.55, …, 0.95}。如当 OKS=0.5 时,则 $AP^{0.5}$。

3.4　推理速度对比实验

为了验证本文基于关键点回归的推理速度的有效性。在 COCO val2017 验证集上进行对比实验,实验在同一设备进行(RTX2060)。由表 1 可以看出,本文的推理时间明显快于 DEKR(disentangled keypoint regression)-W32[24], HRNet-W48, HigherH RNet-W32[25]等方法。以上 3 种方法均是以 HRNet 为基础网络,其在网络推理过程中,特征图保持了较高的分辨率,导致网络的推理速度较低。本文的推理速度虽相比于 KAPAO-S 慢了 1.6 ms,分析是因为本文采用的 SDR-NMS 算法进行了两次遍历,使得本文的后处理时间相较于 KAPAO-S 慢了 1.1 ms,且在网络的输出部分由于加入了并行分支使得网络参数量增加,导致推理速度减慢。但是本文的 AP 分数相比于 KAPAO-S 提高了 4.8%。相比于 KAPAO-M 的 AP 分数提高 0.1%,且速度是其 1.7 倍。在后处理过程中,本文摒弃热图回归,后处理时间相比 HRNet-W48 等算法高了 15~25 倍。

表 1　推理速度对比实验

Table 1　Comparison experiment of inference speed

模型	输入大小	参数/M	准确率/%	召回率/%	前置处理时间/ms	后处理时间/ms	总时间/ms
DEKR-W32	512×512	29.6	62.4	69.6	62.6	34.9	97.5
HRNet-W48	256×256	63.6	74.2	79.5	81.2	53.9	135.1
HigherHRNet-W32	512×512	28.6	63.6	69.0	46.1	50.1	96.2
KAPAO-M	1 280×1 280	35.8	68.5	75.5	30.7	2.9	33.5
KAPAO-S	640×640	12.5	63.8	70.2	14.7	2.8	17.5
本文	640×640	13.9	68.6	79.8	15.2	3.9	19.1

3.5　不同算法的对比实验

同时在 COCO test-dev 测试集中,证明了本文算法的精确率。分别选用先进的方法进行对比试验。由表 2 可知,对比实验分为基于热图回归的和基于关键点回归的

两组。其中表 2 上半部分为基于热图的方法,下半部分为基于关键点回归的方法。

表 2　不同算法的对比实验

Table 2　Comparative experiments of different algorithms

(%)

模型	准确率	$AP^{0.5}$	$AP^{0.75}$	召回率
CPN	72.1	91.4	80.0	78.5
HRNet-W32	73.4	89.5	80.7	78.9
HigherHRNet-W32	66.4	87.5	72.8	70.3
Ref.[5]	72.6	89.5	80.0	—
OpenPose	61.8	84.9	67.5	66.5
CenterNet	63.0	86.8	69.6	—
SPM	66.9	88.5	72.9	—
KAPAO-M	68.5	89.7	76.0	76.3
Mask R-CNN +RLE	67.1	86.7	72.6	—
PRTR-W48	66.2	85.9	72.1	72.2
本文	68.6	90.1	79.2	79.8

本文方法显著优于其他基于关键点回归的方法。相比于 KAPAO-M 在参数量少 1 倍的情况下,AP 分数提高了 0.1%,相比于热图回归到方法,本方法超过了 OpenPose, HigherHRNet-W32 等方法, $AP^{0.5}$ 分数为 90.1%,仅比最优方法 CPN 的 $AP^{0.5}$ 分数低了 1.4%。虽然 AP 分数距最优的方法 HRNet-W32 还有一定距离,但是由表 1 可知,本文的实时性表现是 HRNet 等基于热图回归的方法无法比拟的,本文的速度相比于 HRNet 提高了 7 倍。以上结果表明本文算法在关键点检测方面的有效性。

3.6　消融实验

在 COCO val2017 验证集上进行消融实验,实验结果如表 3 所示。其中“Π”表示使用该模块,由实验结果表明数据增强方法扩充了数据集,与本文方法直接耦合,通过对生成的多样的姿态的关键点的训练,使得 AP 分数提高了 1.1%。BiFPN 相较于 PANet 通过对特征图的加权融合,去除冗余特征图的干扰,使得 AP 分数提高了 0.3%。A-BiFPN 在 BiFPN 的基础上在不同特征图融合部分加入注意力机制,增强了对不同语义信息特征图的融合能力,使得 AP 分数提高了 1.6%。ADC 通过不同扩张率的扩张卷积自适应的获得不同尺寸的感受野区域,使得 AP 分数提高了 0.8%。在以上基础上,设计 SDR-NMS 算法。在 AP 分数提高了 1.3%的同时,AR 分数提高了 3.7%,分析由于 SDR-NMS 使用置信度抑制方法提高关键点对象的召回率,使用重定位方法提高了关键点的定位精确度。

表 3　不同模块的消融实验

Table 3　Ablation experiments of different modules

PoseTrans	BiFPN	A-BiFPN	ADC	SDR-NMS	准确率/%	召回率/%
					63.8	70.2
Π					64.9	71.3
Π	Π				65.2	72.8
Π		Π			66.5	74.9
Π		Π	Π		67.3	76.1
Π		Π	Π	Π	68.6	79.8

3.7　可视化实验

本文在 COCO test-dev 测试集进行可视化实验证本文在多尺度,多人遮挡,小目标,罕见姿态等场景下的有效性和泛化性。并将可视化结果与原始 KAPAO-S 网络进行对比,如图 10 所示,分别为遮挡,多人小目标,多尺度,多人遮挡,罕见姿态的检测结果。

从图 10 中可以看出,本文算法与 KAPAO-S 模型对特征较强的人脸关键点和姿态对象均具有较好的表现。但 KAPAO-S 模型在遮挡,多尺度等具有挑战性的场景下对局部较弱的关键点,如脚踝,臀部等关键点,表现较弱。图 10(a)表明在遮挡的场景下,本文算法能够对关键点进行正确的空间建模,更好的匹配人体姿态。图 10(b)表明在多人和小目标场景下,能够检测出原始算法未检测出的小目标。图 10(c)表明本文对于多尺度同样具有较好的鲁棒性,能够检测出原网络未检测出的臀部关键点。图 10(d)表明本文在多人密集场景下,能够检测出更多的人体关键点,具有较强的检测能力。图 10(e)表明本文算法对罕见姿态的泛化性更好,能够检测出局部特征较弱的人体关键点。

4　结　论

本文提出一种改进的 KAPAO 网络,有效提高了关键点检测在具有挑战性的场景下对局部特征较弱的关键点的检测精确度。使用 PoseTrans 数据增强方法提高模型的泛化性,增强对罕见姿态的检测能力;同时提出一种注意力融合机制的 BiFPN 模块,在特征融合部分引入空间注意力,增强特征融合能力;并设计自适应扩张卷积,自适应获得特征图不同的感受野,从而克服多尺度问题。并设计一种 SDR-NMS 方法,借助该方法提高关键点检测精确度。在 COCO 数据集上进行推理速度和精确度实验,验证了本算的有效性和实时性。

图 10　可视化实验结果

Fig. 10　Visualize the results of experiments

参考文献

[1] 余金锁,卢先领. 基于分割注意力的特征融合 CNN-Bi-LSTM 人体行为识别算法[J]. 电子测量与仪器学报, 2022,36(2):89-95.
YU J S, LU X L. Human action recognition algorithm of feature fusion CNN-Bi-LSTM based on split-attention[J]. Journal of Electronic Measurement and Instrumentation, 2022, 36(2): 89-95.

[2] 陈姝琪,曹江涛,赵挺,等. 基于关节点数据的双人交互行为识别[J]. 电子测量与仪器学报, 2020, 34(6): 124-130.
CHEN SH Q, CAO J T, ZHAO T, et al. Two-person interaction behavior recognition based on joint data[J]. Journal of Electronic Measurement and Instrumentation, 2020, 34(6): 124-130.

[3] CAO Z, SIMON T, WEI S, et al. Realtime multi-person 2d pose estimation using part affinity fields [C]. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA: IEEE Press, 2017: 1302-1310.

[4] CHEN Y L, WANG Z C, PENG Y X, et al. Cascaded pyramid network for multi-person pose estimation[C]. Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA: IEEE Press, 2018: 7103-7112.

[5] WANG J D, SUN K, CHENG T H, et al. Deep high-resolution representation learning for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(10): 3349-3364.

[6] YU C Q, XIAO B, GAO C X, et al. Lite-HRNet: A lightweight high-resolution network[C]. Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA: IEEE Press, 2021: 10435-10445.

[7] MA N N, ZANG X Y, ZHENG H T, et al. Shufflenet v2: Practical guidelines for efficient cnn architecture design[C]. Proceedings of European Conference on Computer Vision (ECCV), 2018: 122-138.

[8] 钟宝荣,吴夏灵. 基于高分辨率网络的轻量型人体姿态估计研究[J]. 计算机工程, 2023, 49(4): 226-232.
ZHONG B R, WU X L. Research on lightweight human pose estimation based on high resolution network[J]. Computer Engineering, 2023, 49(4): 226-232.

[9] HAN K, WANG Y H, TIAN Q, et al. GhostNet: More features from cheap operations[C]. Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE Press, 2020: 1577-1586.

[10] 马皖宜,张德平. 基于多谱注意力高分辨率网络的人体姿态估计[J]. 计算机辅助设计与图形学学报, 2022, 34(8): 1283-1292.
MA W Y, ZHANG D P. Human pose estimation based on multi-spectral attention and high resolution network[J]. Journal of Computer-Aided Design & Computer Graphics,

2022, 34(8): 1283-1292.
[11] TOSHEV A, SZEGEDY C. DeepPose: Human pose estimation via deep neural networks[C]. Proceed ings of 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA: IEEE Press, 2014: 1653-1660.
[12] ZHOU X Y, WANG D Q, KRAHENBUHL P. Objects as points[J]. arXiv preprint arXiv: 1904. 07850, 2019.
[13] LI J F, BIAN S Y, ZENG A L, et al. Human pose regression with residual log-likelihood estimation [C]. Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada: IEEE Press, 2021: 11005-11014.
[14] NIE X C, FENG J, ZHANG J, et al. Single-stage multi-person pose machines[C]. Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea: IEEE Press, 2019:6950-6959.
[15] LI K, WANG S J, ZHANG X, et al. Pose recognition with cascade transformers [C]. Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA: IEEE Press, 2021: 1944-1953.
[16] MCNALLY W, VATS K, WONG A, et al. Rethinking keypoint representations: Modeling keypoints and poses as objects for multi-person human pose estimation[C]. Proceedings of European Conference on Computer Vision (ECCV), 2022: 37-54.
[17] JIANG W T, JIN S, LIU W T, et al. PoseTrans: A simple yet effective pose transformation augmentation for human pose estimation[C]. Proceedings of the European Conference on Computer Vision (ECCV), 2022: 643-659.
[18] LIU S, QI L, QIN H F, et al. Path aggregation network for instance segmentation [C]. Proceeding s of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Salt Lake City, UT, USA: IEEE Press, 2018: 8759-8768.
[19] TAN M X, PANG R M, LE Q V. EfficientDet: Scalable and efficient object detection[C]. Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA: IEEE Press, 2020: 10778-10787.
[20] LIN T Y, DOLLAR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA: IEEE Press, 2017: 936-944.
[21] BODLA N, SINGH B, CHELLAPPA R, et al. Soft-NMS-improving object detection with one line of code [C]. Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, New York: IEEE Press, 2017: 5562-5570.
[22] GOYAL P, DOLLÁR P, GIRSHICK R, et al. Accurate, large minibatch sgd: Training imagenet in 1 hour[J]. arXiv preprint arXiv:1706. 02677, 2017.
[23] LOSHCHILOV I, HUTTER F. SGDR: Stochastic gradient descent with warm restarts[C]. Proceedings of International Conference on Learning Representations, Toulon, France, 2017.
[24] GENG Z G, SUN K, XIAO B, et al. Bottom-up human pose estimation via disentangled keypoint regression[C]. Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA: IEEE Press,2021: 14671-14681.
[25] CHENG B W, XIAO B, WANG J D, et al. HighterHRNet: Scale-aware representation learning for bottom-up human pose estimation [C]. Proceedings of 2020IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA: IEEE Press, 2020: 5385-5394.

作者简介

赵普,2020 年于南阳理工学院获得学士学位,现为河北工业大学电子信息工程学院在读研究生,主要研究方向为姿态估计、跌倒识别。
E-mail: zp_dling@ 163. com

Zhao Pu received his B. Sc. degree from Nanyang Institute of Technology in 2020. He is now a M. Sc. candidate in the School of Electronics Information Engineering at Hebei University of Technology. His main research interests include human pose estimation and fall detection

武一(通信作者),现为河北工业大学电子信息工程学院教授,主要研究方向为智能系统控制与应用。
E-mail: wuyihbgydx@ 163. com

Wu Yi (Corresponding author) is now a professor in the School of Electronic Information Engineering at Hebei University of Technology. Her main research interests include system control and application.