

基于改进高分辨率网络的人体姿态估计

刘洁¹, 陈志¹, 岳文静²

(1. 南京邮电大学 计算机学院; 2. 南京邮电大学 通信与信息工程学院, 江苏 南京 210003)

摘要: 为实现更精准的人体关键点定位, 以高分辨率检测网络(HRNet)为基线引入瀑布式空洞空间卷积模块与Transformer的人体姿态估计模型和算法。首先, 构建瀑布式空洞空间卷积模块替换HRNet的第4阶段, 减少不同尺度特征相互融合导致参数量过大的问题, 并更高效地提取多尺度特征; 其次, 引入基于自注意力机制的Transformer对提取的高层特征进行处理, 通过捕获全局空间中关键点的非局部交互关系以获取全局信息实现特征增强。实验表明, 当输入图像分辨率为256×192时, 所提模型相较于HRNet-W32和HRNet-W48基线模型, 在参数量下降的情况下AP分别提升2.4%、2.3%。

关键词: 人体姿态估计; 高分辨率网络; 瀑布式空洞卷积; 注意力机制; 多尺度

DOI: 10.11907/rjdk.241182

中图分类号: TP391

文献标识码: A

开放科学(资源服务)标识码(OSID):

文章编号: 1672-7800(2024)006-0136-07



Human Pose Estimation Based on Improved High-Resolution Network

LIU Jie¹, CHEN Zhi¹, YUE Wenjing²

(1. School of Computer Science, Nanjing University of Posts and Telecommunications; 2. School of Communications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: To achieve more accurate positioning of human body key points, a human pose estimation model and algorithm are introduced based on a high-resolution detection network (HRNet) with a waterfall shaped cavity spatial convolution module and Transformer. Firstly, a waterfall like hollow space convolution module is constructed to replace the fourth stage of HRNet, reducing the problem of large parameter quantities caused by the fusion of features at different scales and extracting multi-scale features more efficiently; Then, a Transformer based on self attention mechanism is introduced to process the extracted high-level features, and feature enhancement is achieved by capturing the non local interaction relationships of key points in the global space to obtain global information. The experiment shows that when the input image resolution is 256×192 , the proposed model improves AP by 2.4% and 2.3% respectively compared to the HRNet-W32 and HRNet-W48 baseline models with a decrease in parameter count.

Key Words: human pose estimation; high-resolution network; waterfall dilated convolution; attention mechanism; multi-scale

0 引言

人体姿态估计作为计算机视觉领域的关键研究方向一直以来备受关注, 具体的研究任务是从图片或视频中检测人体的关节点并重构人体骨架。目前, 人体姿态估计在众多领域得到了广泛应用, 包括动作识别^[1]、运动捕捉^[2]和身份识别^[3]等, 尤其是身份识别中用到的人体关键点, 原理是基于人体骨架关键点对人物进行建模, 具有计算速

度快、效果良好的优点。

目前, 随着卷积神经网络和循环神经网络的发展, 基于深度学习的人体姿态估计方法已取得了显著进展, 在模型精度方面远超传统方法^[4-6]。Toshev等^[7]提出一种神经网络DeepPose, 第一次将深度学习方法应用在人体姿态估计领域, 利用卷积神经网络从图像中提取特征和关键点坐标回归, 将传统方法的图像处理问题转换为卷积神经网络中的坐标回归问题, 但仅使用坐标回归方法的效果较差。

根据人体关键点检测顺序, 检测方法可分为自上而下

收稿日期: 2024-02-29

基金项目: 江苏省重点研发计划(社会发展)项目(BE2019739); 中兴通讯产学研合作基金项目(2021H2ZTE05-01)

作者简介: 刘洁(1998-), 男, 南京邮电大学计算机学院硕士研究生, 研究方向为计算机视觉; 陈志(1978-), 男, 博士, 南京邮电大学计算机学院教授, 研究方向为软件工程、无线传感网、计算机视觉; 岳文静(1982-), 女, 博士, 南京邮电大学通信与信息工程学院副教授, 研究方向为信号处理、边缘计算、计算机视觉。本文通讯作者: 陈志。

(Top-Down)、自下而上(Down-Top)两种。其中,自上而下(Top-Down)的关键点检测顺序是先进行目标检测,然后寻找图片中的人物并截取,随后将其输入网络进行关键点检测;自下而上(Down-Top)的关键点检测顺序是从图片中检测关键点,然后采取匹配或聚类的方式组合检测出的关键点^[8-10]。Chen等^[11]提出一种自上而下的级联金字塔网络(Cascaded Pyramid Network, CPN),该网络主要分为两个重要部分:①ClobalNet主要检测容易识别的关键点;②RefineNet处理较难检测的关键点,例如解决复杂场景下的关键点检测、输入图片尺寸太小和人体关键点部分遮挡等。Xiao等^[12]提出一种简单基线模型 Simple Baselines,使用ResNet网络作为骨干网络进行特征提取,网络尾部采用反卷积恢复分辨率,但在处理时先降低分辨率,导致关键点定位精度不高^[13]。Sun等^[14]提出一种高分辨率网络 HRNet,核心思想是保留高分辨率的信息,而不是将其降采样为较低分辨率。HRNet 允许网络在多个尺度上学习特征,在不同分辨率上构建多个并行的卷积分支,并将这些分支特征融合在一起,但多分辨率分支和信息融合机制增加了网络的计算复杂度和参数量,在复杂场景下无法很好地预测关键点位置的坐标,因此会影响检测效果。Wang等^[15]提出一种减少网络深度的模型 Small HRNet,但模型预测精度较差。

近年来,不少研究者作了较为全面的综述研究。吴佳豪等^[19]提出一种堆叠沙漏网络的方法获取更丰富的特征表示,但网络计算复杂度高。许伟^[20]提出一种基于热力图缩小高斯内核的人体姿态估计方法。陶杭宇等^[21]使用较小的卷积进行多尺度特征融合,虽然提高了网络精度,但无法保持高分辨率特征表示^[16-18]。

针对以上问题,本文基于 HRNet-W32 和 HRNet-W48 的高分辨率网络结构,研究和设计了引入瀑布式空洞空间卷积模块与注意力机制的人体姿态估计模型和算法,通过构建瀑布式空洞空间卷积模块替换 HRNet 的第4阶段,以扩大网络感受视野,并进行高效的多尺度特征提取,从而实现特征增强的目的。此外,在最后阶段采用基于自注意力机制的 Transformer 处理高层特征获取人体关键点的空间关系,以提升模型预测关键点位置的精度。

1 相关工作

高分辨率网络(High-Resolution Network)是一种可全程保持高分辨率特征的新型网络结构。传统卷积神经网络(CNN)在处理图像时通常会采用下采样操作,将导致特征图的分辨率降低,而 HRNet 能在网络内的多个层级同时维持高分辨率表示。HRNet 网络主体包括4个阶段,每个阶段均由并行的高分辨率子网连接而成,起始阶段由两个 3×3 的卷积核对图像进行预处理,将输入图像的分辨率下采样至 $1/4$ 。第1阶段是一个高分辨率子网,第2、第3和第

4阶段由多分辨率模块构成,每个部分的形成过程是将前一阶段最低分辨率的 $1/2$ 作为新的旁支,并行整合到当前分支,如图1所示。HRNet 通过并行多分支结构使新分支一直保留前一阶段的高分辨率信息并进行交叉融合,以提高网络整体的特征表示。

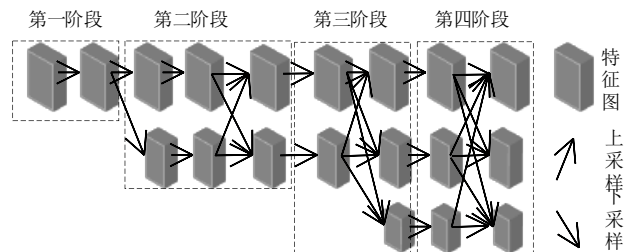


Fig. 1 HRNet model

图1 HRNet模型

并行多尺度融合多用于处理图像中不同尺度特征信息的融合。通过并行处理不同分辨率的特征,网络能更全面捕捉图像中的细粒度信息,提高网络对不同尺度特征的感知能力。在并行多尺度特征融合时,每个阶段分支网络会产生大量参数,从而导致网络计算复杂度增加,因此在高分辨率网络 HRNet 中进行多尺度特征融合时会设置一些 1×1 的卷积核处理低分辨率特征,然后进行最近邻上采样。其中,对高分辨率特征的下采样通过设置 3×3 的卷积核来处理,当分辨率相同时不进行任何操作,如图2所示。

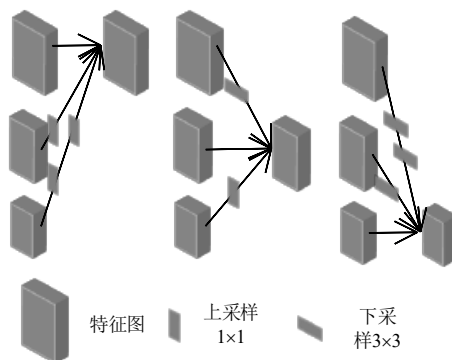


Fig. 2 Feature fusion with different resolutions

图2 不同分辨率特征融合

特征提取的每个阶段包含了一系列高、中、低分辨率的分支,不同阶段的横向连接有所不同,假设高分辨率分支输出为 H ,中分辨率分支的输出为 M ,低分辨率分支的输出为 L ,用 $Upsample$ 表示上采样操作, $Conv_{1 \times 1}$ 、 $Conv_{3 \times 3}$ 表示卷积操作,则横向连接可表示为:

$$H' = H + Upsample(Conv_{1 \times 1}(M)) + Upsample(Conv_{1 \times 1}(L)) \quad (1)$$

$$M' = M + Upsample(Conv_{1 \times 1}(L)) + Conv_{3 \times 3}(H) \quad (2)$$

$$L' = L + Conv_{3 \times 3}(H) + Conv_{3 \times 3}(M) \quad (3)$$

Transformer 是由具有自注意力机制的编码器和解码器组成,最初用于语言建模和机器翻译等任务,目前已被应用在计算机视觉、音频信号处理等领域。本文使用自注

注意力机制对提取的高层特征进行处理,实际上是让机器注意到整个输入中不同部分之间的相关性,处理后会得到每个部分感兴趣部分^[22]。具体的,输入由 Q, K, V 矩阵组成,假设第 t 帧的个体特征为 $X^t \in R^{T \times N \times D}$, 相应的矩阵计算公式为:

$$Q^t = X^t W^{tq}, K^t = X^t W^{tk}, V^t = X^t W^{tv} \quad (4)$$

然后,将 Q 与 K 的转置相乘并除以 $\sqrt{D}$, 使用 softmax 层进行归一化以获得注意力分数。

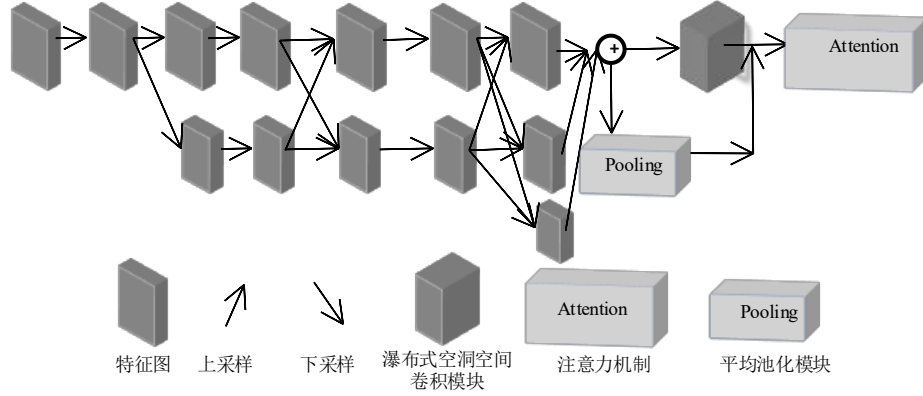


Fig. 3 Improved overall network structure

图3 改进后的网络整体结构

HRNet网络通过并行融合多尺度特征获得高分辨率表示,但后续每个阶段需要生成一个新的分支对多尺度特征进行融合,特别在最后阶段将产生大量参数。针对该问题,本文引入瀑布式空洞空间卷积模块代替HRNet网络的第4阶段,不仅可获得多尺度特征的表现,还扩大了网络感受视野。

改进后的HRNet网络只包含3个阶段:①起始阶段输入图像经过两个 3×3 的卷积、两个归一化层和ReLU激活函数初步提取特征,将输入图像的分辨率下采样至 $1/4$,然后在第1阶段利用残差结构加深网络层数提取特征,主要由一个高分辨率子网构成,并使用4个Bottleneck模块进行特征提取;②第2阶段包含两个多分辨率模块;③第3阶段包含3个多分辨率模块,对每个模块的处理都包含4个BasicBlock模块以加深网络层数进行特征提取,每个BasicBlock模块包含两个 3×3 的卷积、两个归一化层和ReLU激活函数。

根据上述分析,第4阶段将使用 3×3 卷积操作增加网络参数数量和计算复杂度,而且HRNet网络的残差模块通常采用 3×3 的普通卷积,在一定程度上限制了网络的感受视野,导致模型无法检测一些小尺度的特征。瀑布式空洞空间卷积模块中每层包含不同尺度大小的卷积核,较小尺寸的卷积核有助于聚焦图像的细粒度信息,从而更好地检测关键点,且网络参数数量更少;较大尺寸的卷积核能更可靠地捕捉较大目标信息。通过该方式可在降低计算复杂度和减少参数数量的同时,有效提取深层、浅层的图像特征,意味着网络能以更高效的方式学习图像的多层次表示,从而在处理复杂任务时取得良好的性能,无需过多的计算资源

2 高分辨率人体姿态估计模型优化

2.1 高分辨率特征提取

本文模型以高分辨率网络HRNet为基线,改进后的网络整体结构如图3所示。

和参数设置。

此外,本文引入瀑布式空洞空间卷积模块代替高分辨率网络HRNet的第4阶段,对HRNet网络第3阶段输出的高分辨率特征直接使用瀑布式空洞空间卷积进行多尺度特征提取,无需再进行下采样和上采样操作,因为使用下采样和上采样操作会降低图像分辨率,如图4所示。

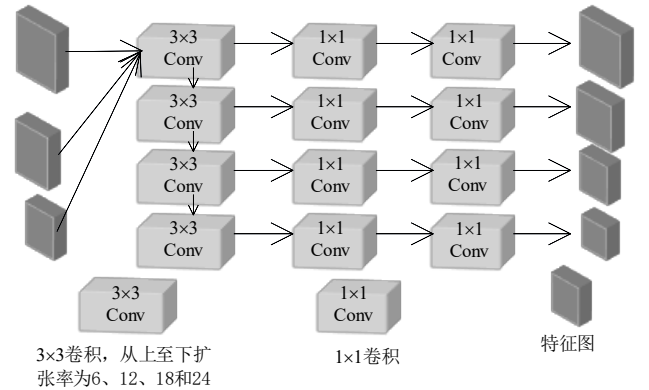


Fig. 4 Waterfall type hollow space convolution module

图4 瀑布式空洞空间卷积模块

由图4可见,该模块拥有空洞空间卷积金字塔(Atrous Spatial Pyramid Pooling, ASPP)模块的特点,特征提取时结合空洞卷积能保持较大的感受视野,通过串联方法与 1×1 卷积操作可减少模型参数数量^[23]。空洞卷积的主要目的是以膨胀的方式扩大网络对图像的感受野,从而避免过多下采样操作,以生成适用于处理多尺度信息的框架,使网络能更好地捕捉图像中不同尺度的特征,在复杂场景下可提供更全面的信息。在一维卷积情况下的计算公式为:

$$Att^t = \text{softmax} \left(\frac{Q^t K^{t^T}}{\sqrt{D}} \right) V^t \quad (5)$$

$$h[i] = \sum_{x=1}^x f[i+rx] * w[l] \quad (6)$$

式中: $h[i]$ 为输出; $f[i]$ 为输入; r 为扩张率; $w[l]$ 为滤波器。

瀑布式空洞空间卷积模块在处理过程中会产生4个不同分辨率的特征,输入特征不会立即使用并行卷积进行处理,而是以串联卷积方式通过不同的采样率生成一个新分支,从而实现瀑布流,且每个分支具有不同的感受视野。因此采样率的设置不能太小或太大,否则将无法提取关键的语义信息。

本文首先通过4层串联的空洞卷积进行采样,空洞系数分别为6、12、18和24,以瀑布方式进行排列;其次对生成的每个新分支进行两次 1×1 的卷积操作,以提取图像细粒度信息;最后将经过平均池化后的原始特征与4个分支的输出特征相互融合,以完成多尺度特征表示。

2.2 高分辨率特征注意力

卷积神经网络虽然能从图像中提取特征信息,但无法表示全局特征间的空间关系。在人体姿态估计任务中,卷积神经网络通过相互叠加可提取出小部位关键点,例如耳朵和眼睛,但卷积操作无法捕捉这些部位之间的位置关系,无法考虑视野外的上下文信息。因此,在处理多特征之间的相互关系时,需要结合注意力机制以更好地捕捉全局特征信息之间的交互关系。

为此,本文引用基于自注意力机制的Transformer模块对人体关键点的相对关系进行建模,根据式(5)的自注意力机制计算全局特征的注意力权重,使每个输入特征向量都能利用其他输入的特征信息进行编码。由于标准Transformer的自注意力机制会计算全局信息,应用到人体姿态估计任务中会产生较大的计算量,因此通过添加掩膜操作优化注意层,以降低网络计算量。

具体为,本文首先假设网络输入图像大小为 $X \in \mathbb{R}^{3 \times H \times W}$,经过CNN主干网络特征提取后的输出特征大小为 $X' \in \mathbb{R}^{d \times H \times W}$;然后将图像特征转换为 $X'' \in \mathbb{R}^{L \times d}$,即 $L = H \times W$,表示为 L 个 d 维向量;最后将向量输入Transformer中使用PE函数对特征图的像素点进行位置编码,然后融合特征序列和位置编码信息表示输入,具体计算方法为:

$$PE(pos, 2i) = \sin(pos/10000^{2i/D}) \quad (7)$$

$$PE(pos, 2i+1) = \cos(pos/10000^{2i/D}) \quad (8)$$

式中:PE表示位置嵌入; pos 表示像素点的位置信息; D 表示维度; $2i, 2i+1$ 表示该维度处于偶数位置和奇数位置。

在特征经过多头注意层时,通过掩膜操作 $M(*)$ 选择最大的 h 个权重。具体为:通过注意力层计算得到注意力分数矩阵 W 后进行 $softmax$ 操作;然后选择 W 的每一行最大的 h 个元素进行归一化。上述操作中,对每一行选择最大值的元素时会记下所在矩阵的位置,假设矩阵某一行第

h 大值为 y_i , i 代表第 i 行,如果该行后面的某个值大于 y_i 则记录当前位置,对注意力分值矩阵 P 进行掩膜操作。

$$M(W, h)_{ij} = \begin{cases} W_{ij}, & W_{ij} - y_i \geq 0 \\ 0, & W_{ij} - y_i \leq 0 \end{cases} \quad (9)$$

式中: i, j 代表坐标位置;通过掩膜操作获得 W 上 h 个最大注意力权重;通过显式选择保留重要信息; h 还可设置为更小的值,以简化模型、减小计算量,提供更精细的控制,使模型在注意力分配上更灵活。

3 基于优化高分辨率网络的人体姿态估计算法设计

3.1 高分辨率特征提取的算法设计

第一步:模型输入图片的尺寸大小为 256×192 ,图片首先经过两个 3×3 的卷积进行初步特征提取;然后将分辨率下采样 $1/4$,经过HRNet的前3阶段进行多尺度特征提取,得到 $16 \times 32 \times 64 \times 48$ 的高分辨率特征图。

第二步:将特征图进行平均池化操作,包括一个平均自适应池化层、一个 1×1 的卷积、一个归一化层和ReLU激活函数。平均自适应池化层无需指定卷积大小,通过设定输出大小即可完成池化操作(本文设置为 1×1),平均自适应池化通过动态调整池化窗口大小,可更好地保留输入的重要信息,不会因输入尺寸变化而导致信息损失,最后使用 1×1 的卷积操作降维特征。

第三步:将第一步获取的特征图输入瀑布式空洞空间卷积模块,在处理过程中依次经过空洞系数为6、12、18和24的卷积操作以获取丰富的上下文信息,然后与上一步输出的特征图通过concatenate操作进行连接,得到高分辨率特征图。

算法1 改进后的模型训练算法

输入:训练集 $Train(x)$ 、数量 H 、验证集 $Val(x)$ 、总轮数 $Epoch$ 、批处理大小 B 。

输出:模型 $F(x)$ 。

1. 初始化网络模型随机权重

2. FOR $i := 1$ to $Epoch$ DO

3. FOR $j := 1$ to H/B DO

4. 从 $Train(x)$ 中选择批量大小为 B 的样本 $(x_1, x_2, \dots, x_b)$, $1 \leq b \leq B$

5. 分支网络进行特征提取 $F'(x) \rightarrow x'_k$

6. 特征融合 $FUSION(x'_k) \rightarrow x''_k$

7. 平均池化操作 $AvgPooling(x''_k) \rightarrow x_p$

8. 输入瀑布式空洞空间卷积模块 $F''(x''_k) \rightarrow x'''_k$

9. 特征融合 $FUSION(x'''_k) \rightarrow x_f$

10. concatenate(x_p, x_f)

11. END FOR

12. END FOR

3.2 基于优化高分辨率网络的人体姿态估计算法整体设计

第一步:模型输入图片大小为 256×192 ,通过优化后的骨干网络提取特征,得到大小为 $3072 \times 16 \times 64$ 的高分辨率

特征图。

第二步:将图像输入优化后的 Transformer 模块,以建模关键点间的空间相关性。首先进行位置编码,将特征图展平得到大小为 $d \times l$ 的序列;然后使用 sin-cos 规则对特征图进行位置编码;最后将所得特征输入 Transformer 中捕获全局空间中关键点的非局部交互关系。

第三步:首先比较预测的人体关键点位置与数据集中的真实标签的差异;然后使用均方误差计算损失值,通过更新神经网络参数以减少损失值;最后经过多次迭代,逐渐优化网络参数并得到本文最佳模型。

算法 2 改进后的整体模型训练算法

输入:训练集 $Train(x)$ 和数量 H 、验证集 $Val(x)$ 、总轮数 $Epoch$ 、批处理大小 B 。

输出:模型 $T(x)$ 。

1. 初始化网络模型随机权重
2. FOR $i := 1$ to Epoch DO
3. FOR $j := 1$ to H/B DO
4. 从 $Train(x)$ 中选择批量大小为 B 的样本 $(x_1, x_2, \dots, x_b)$, $1 \leq b \leq B$
5. 特征提取 $F(x) \rightarrow x'$
6. 位置编码 $PE(pos, 2i) = \sin(pos/10000^{2i/D})$
7. 利用 $Transformer(Q, K, V) \rightarrow x$ 进行建模
8. 预测关键点位置 $T(x) \rightarrow Y$
9. 使用均方误差计算损失值 $loss$
10. 最小化 $loss$ 并更新模型 $T(x)$ 参数
11. END FOR
12. 使用验证集 $Val(x)$ 测试模型准确率
13. END FOR

4 实验结果与分析

4.1 数据集与评价指标

本文实验训练与测试采用的公共数据集为 COCO2017,由 20 多万张被标记的图片组成,数据集涵盖了 25 万个人体实例,每个实例都标注了 17 个关键点。COCO2017 训练集包含 5 700 张图像、15 万个人体实例,验证集有 5 000 张图像,测试集有 20 000 张图像。本文模型将在 COCO2017 训练集上进行训练,在 COCO2017 验证集和测试集上进行测试评估,评价标准使用 COCO 官方提供的 OKS(Object Keypoint Similarity),如式(10)所示。

$$OKS_p = \frac{\sum_i \exp\left(\frac{-d_{pi}^2}{2S_p^2\sigma_i^2}\right)\delta(V_{pi} > 0)}{\sum_i \delta(V_{pi} > 0)} \quad (10)$$

式中: P 代表图像中某人ID; P_i 代表某个人关键点ID; d_{pi} 代表预测值与真实值之间的欧氏距离; S_p^2 代表某人所占面积; σ_i 表示归一化因子; δ 表示关键点函数且输出只有 0 或 1。

COCO 数据集一般包含平均准确度(Average Preci-

sion, AP)和平均召回率(Average Recall, AR)两个重要指标,可通过计算预测值关键点与真值关键点的相似性得到 AP 和 AR 指标。通常情况下, AP^{50} 和 AP^{75} 分别表示阈值为 0.5 和 0.75 时的准确率,AP 指 OKS 等于 0.5、0.55、 $\dots$ 、0.9、0.95 时的平均精度; AP^M 为中等尺度目标精度; AP^L 为较大尺度目标的精度。同理,AR 也具有相同的计算过程。

4.2 实验环境与参数

本文实验使用 Ubuntu 系统作为运行平台,采用深度学习框架 Pytorch 训练模型,编程语言主要为 Python 且通过 GPU 进行加速,具体实验环境如表 1 所示。

Table 1 Experimental environment configuration

表 1 实验环境配置		
环境	配置	配置参数
硬件环境	操作系统	Ubuntu 20.04
	CPU 型号	Intel i9-10900K
	GPU 型号	NVIDIA GeForce RTX 3090
	显存	24 GB
软件环境	CUDA 版本	11.4
	编程语言	Python
	深度学习框架	Pytorch

模型训练时对图像进行预处理,固定输入图像大小为 256×192 ,模型参数优化器为 Adam,训练批次大小为 16,网络训练一共 240 个周期,学习率开始设置为 $1e-4$ 并使用余弦退火的学习率下降方式,减小到 0.000 01 不再衰减^[24]。

4.3 实验比较

为了验证本文方法的有效性,在 COCO2017 验证集上,分别进行基于 HRNet-W32 和 HRNet-W48 网络改进的模型实验,并与其他方法进行比较,结果如表 2 所示。数据表明,本文方法相较于其他人体姿态估计方法,在关键点预测上的估计精度较高。

同时,在 COCO2017 测试集上将本文方法与其他方法进行比较,输入图像尺寸大小统一为 384×288 ,实验结果如表 3 所示。由此可知,本文所提模型在估计精度上相较于其他方法也存在一定程度的提升。

4.4 消融实验

为验证所提模型各部分对性能的影响,在 COCO2017 验证集上对本文方法进行消融实验。模型输入图像固定大小为 256×192 ,训练批次大小为 16,学习率与训练周期均保持一致,实验结果如表 4 所示。其中,B1 表示 HRNet-W32 网络经过剪枝第 4 阶段后的模块;B2 表示瀑布式空洞空间卷积模块;B3 表示基于自注意力机制的 Transformer 模块。

由表 4 可知,HRNet-W32 剪枝网络与注意力机制模块相结合后,参数仅为基线模型的 29.8% 且平均精度 AP 提高 0.8%,证明了引入自注意力机制模块可充分利用上下文信息处理多特征之间的相互关系,能更好地预测关键位置;在 HRNet-W32 剪枝网络与注意力机制模块结合的基础上加入瀑布式空洞空间卷积模块,参数仅为基线模型的 60.7% 且平均精度 AP 提高 2.4%,虽然相较于未加入瀑布

Table 2 Comparison of experimental results on the COCO2017 validation dataset

表 2 COCO2017 验证集实验结果比较										
模型	Backbone	Inputsize	Params/M	GFLOPs	AP	AP ⁵⁰	AP ⁷⁵	AP ^M	AP ^L	AR
CPN	ResNet-50	256×192	27.0	6.20	68.6	—	—	—	—	—
CPN+OHKM	ResNet-50	256×192	27.0	6.20	69.4	—	—	—	—	—
SimpleBaseline	ResNet-50	256×192	34.0	8.90	70.4	88.6	78.3	67.1	77.2	76.3
SimpleBaseline	ResNet-101	256×192	53.0	12.4	71.4	89.3	79.3	68.1	78.1	77.1
文献[25]	HRNet-W32	256×192	17.5	4.60	72.3	88.1	79.5	68.7	79.1	77.9
文献[26]	HRNet-W32	256×192	30.7	8.09	74.8	—	—	71.2	81.7	—
HRNet	HRNet-W32	256×192	28.5	7.10	73.4	89.5	80.7	70.2	80.1	78.9
	HRNet-W48	256×192	63.6	14.6	75.1	90.6	82.2	71.5	81.8	80.4
our	HRNet-W32	256×192	17.3	11.2	75.8	92.5	82.7	73.1	80.0	78.6
	HRNet-W48	256×192	27.5	17.8	77.4	93.6	84.8	74.7	82.0	80.1

Table 3 Comparison of experimental results on the COCO2017 test dataset

表 3 COCO2017 测试集实验结果比较							
模型	Backbone	Inputsize	Params	GFLOPs	AP	AP ^M	AP ^L
CPN	ResNet-50	384×288	—	—	72.1	68.7	77.2
CPN+OHKM	ResNet-50	384×288	—	—	73.0	69.5	78.1
SimpleBaseline	ResNet-152	384×288	68.6M	15.7	73.7	70.3	80.0
文献[17]	HRNet-W32	384×288	30.7M	18.2	75.3	71.8	81.3
HRNet	HRNet-W32	384×288	28.5M	16.0	74.9	71.3	80.9
	HRNet-W48	384×288	63.6M	32.9	75.5	71.9	81.5
our	HRNet-W32	384×288	17.3M	17.5	75.2	72.6	79.8
	HRNet-W48	384×288	27.5M	23.8	76.3	74.1	81.7

Table 4 Results of ablation experiments

表 4 消融实验结果				
B1	B2	B3	Params	AP
Baseline(HRNet-W32)			28.5	73.4
√	×	√	8.5	74.2
√	√	√	17.3	75.8

式空洞空间卷积模块的模型参数上升有所,但平均精度提高 1.6%,证明了引入瀑布式空洞空间卷积模块可加强多尺度特征的提取且相较于基线模型减少了更多参数量。

4.4 可视化结果分析

图 5 为本文模型在 COCO 数据集中部分图像的可视化结果,图中标注点表示预测的人体关键点位置。由此可见,模型在面对低分辨率图像时模型依然表现出色,成功捕捉并预测到了人体关键点位置,即便图像中存在部分遮挡的情况,仍能较好预测人体关键点位置,证明了本文方法在复杂的场景中仍然具有良好的表现。

5 结语

本文以高分辨率检测网络(HRNet)为基线,提出一种



Fig. 5 Visualization results of human pose estimation

图 5 人体姿态估计可视化结果

引入瀑布式空洞空间卷积模块与 Transformer 的人体姿态估计模型,进一步提升了预测关键点位置的精度。瀑布式空洞空间卷积模块的每一层包含不同尺度大小的卷积核,不仅扩大了网络感受视野还能更高效地学习图像的多层次特征表示。对于提取到的高层特征,使用注意力机制对关键点间的空间相关性进行建模,以提升模型预测关键点位置的精度。

在 COCO2017 验证集的结果表明,本文模型相较于 HRNet-W32 和 HRNet-W48 基线模型,既减少了模型参数,又使 AP 分别提高 2.4%、2.3%,取得了良好的效果。后续将进一步优化模型各个阶段,以提升模型精度。

参考文献:

[1] DUAN H, ZHAO Y, CHEN K, et al. Revisiting skeleton-based action recognition [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 2969–2978.
[2] WEI W L, LIN J C, LIU T L, et al. Capturing humans in motion: temporal-attentive 3D human pose and shape estimation from monocular video [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 13211–13220.

- [3] LIAO R, YU S, AN W, et al. A model-based gait recognition method with body pose and human prior knowledge[J]. Pattern Recognition, 2020, 98: 107069.
- [4] ANDRILUKA M, ROTH S, SCHIELE B. Pictorial structures revisited: people detection and articulated pose estimation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2009: 1014–1021.
- [5] LADICKY L, TORR P H S, ZISSERMAN A. Human pose estimation using a joint pixel-wise and part-wise formulation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2013: 3578–3585.
- [6] ZHANG X K, ZHANG R F, LIU Y H. Human pose estimation based on secondary generative confrontation[J]. Advances in Lasers and Optoelectronics, 2020, 57(20): 335–343.
张显坤, 张荣芬, 刘宇红. 基于二次生成对抗的人体姿态估计[J]. 激光与光电子学进展, 2020, 57(20): 335–343.
- [7] TOSHEV A, SZEGEDY C. Deeppose: human pose estimation via deep neural networks[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2014: 1653–1660.
- [8] CAO Z, SIMON T, WEI S E, et al. Realtime multi-person 2D pose estimation using part affinity fields[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017: 7291–7299.
- [9] LI J, WANG C, ZHU H, et al. Crowdpose: efficient crowded scenes pose estimation and a new benchmark[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 10863–10872.
- [10] CHENG B, XIAO B, WANG J, et al. HigherhrNet: scale-aware representation learning for bottom-up human pose estimation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 5386–5395.
- [11] 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, 2018: 7103–7112.
- [12] XIAO B, WU H, WEI Y. Simple baselines for human pose estimation and tracking[C]// Proceedings of the European Conference on Computer Vision, 2018: 466–481.
- [13] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2016: 770–778.
- [14] 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, 2019: 5693–5703.
- [15] 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.
- [16] LU X, ZHANG J H. A survey in human pose estimation methods based on neural networks[J]. Software Guide, 2020, 19(10): 46–49.
卢旭, 张炯华. 基于神经网络的人体姿态估计方法研究[J]. 软件导刊, 2020, 19(10): 46–49.
- [17] MA Z Y, PENG R Y, SUN X H, et al. Review of human pose estimation technology research based on OpenPose[J]. Software Guide, 2022, 21(11): 247–252.
马子越, 彭瑞阳, 孙晓晗, 等. 基于 OpenPose 的人体姿态估计技术研究综述[J]. 软件导刊, 2022, 21(11): 247–252.
- [18] QIAO Y, QU Y. Overview of 2D human pose estimation based on convolution neural network[J]. Application of Electronic Technique, 2021, 47(6): 15–21.
乔逸, 曲毅. 基于卷积神经网络的 2D 人体姿态估计综述[J]. 电子技术应用, 2021, 47(6): 15–21.
- [19] WU J H, ZHOU F, LI L L. Human pose estimation based on stacked hourglass network[J]. Computer System Applications, 2021, 30(10): 295–300.
吴佳豪, 周凤, 李亮亮. 基于堆叠沙漏网络的人体姿态估计[J]. 计算机系统应用, 2021, 30(10): 295–300.
- [20] XU W. Human pose estimation based on deep learning[D]. Chengdu: University of Electronic Science and Technology of China, 2021.
许伟. 基于深度学习的人体姿态估计[D]. 成都: 电子科技大学, 2021.
- [21] TAO H Y, JIANG X Y. Human pose estimation network based on multi-scale bottleneck layer[J]. Software Guide, 2022, 21(12): 21–25.
陶杭宇, 姜晓燕. 基于多尺度瓶颈层的人体姿态估计网络[J]. 软件导刊, 2022, 21(12): 21–25.
- [22] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[DB/OL]. <https://arxiv.org/abs/1706.03762>.
- [23] ZHAO S, FENG Z, CHEN L, et al. DANet: a semantic segmentation network for remote sensing of roads based on Dual-ASPP structure[J]. Electronics, 2023, 12(15): 3243.
- [24] LOSHCHILOV I, HUTTER F S. SGDR: stochastic gradient descent with warm restarts[DB/OL]. <https://arxiv.org/abs/1608.03983>.
- [25] ZHU C T, LI B. Human pose estimation based on high-resolution networks[J]. Journal of South-Central Minzu University: Natural Science Edition, 2023, 42(2): 229–237.
朱翠涛, 李博. 基于高分辨率网络的人体姿态估计[J]. 中南民族大学学报: 自然科学版, 2023, 42(2): 229–237.
- [26] REN H P, WANG W M, WEI D J, et al. Human pose estimation method based on high resolution network[J]. Journal of Graphics, 2021, 42(3): 432–438.
任好盼, 王文明, 危德健, 等. 基于高分辨率网络的人体姿态估计方法[J]. 图学学报, 2021, 42(3): 432–438.

(责任编辑: 刘嘉文)