基于通道切分的人体姿态估计算法

周昆阳[1] 赵梦婷[1] 张海潮[1] 邵叶秦[2]

（1. 南通大学 张謇学院,江苏 南通 226019；2. 南通大学 交通与土木工程学院,江苏 南通）

摘 要：为了提高人体姿态估计的准确率和识别速度，提出一种基于通道切分的人体姿态估计算法Channel-Split Residual Steps Network(Channel-Split RSN)。首先，提出通道切分模块，对切分后的特征通道通过卷积提取特征再融合起来，以获得丰富的特征表示。接着，引入特征增强模块，对特征通道进一步分组，并对不同的分组采取不同的处理策略，以减少特征通道内的相似特征。最后，结合改进的空间注意力机制，提出一种基于特征空间相关性的姿态修正机Context-PRM，得到更加准确的人体姿态估计。在COCO test-dev数据集上的实验结果表明，本文方法达到75.9%的AP和55.36的FPS，并且模型的大小Params(M)仅为18.3。相较于传统的RSN18和传统的RSN50，模型的AP分别提高5%和3.4%，FPS比传统的RSN50快12.08。在更具挑战性的CrowdPose数据集上，本文方法达到66.9%的AP和19.16的FPS，相较于RSN18，AP提高4.6%。有效提高了人体姿态估计的准确率，且模型具有较快的识别速度。本文源代码公开在https://github.com/qdd1234/Channel-Split-RSN。

关键词：Channel-Split RSN；人体姿态估计；通道切分模块；特征增强模块；Context-PRM

中图分类号：TP391.41　TH7　**文献标识码：**A　**国家标准学科分类代码：**510.80

Human Pose Estimation Algorithm based on Channel Splitting

ZHOU Kunyang[1], ZHAO Mengting[1], ZHANG Haichao[1], SHAO Yeqin[2]

（1. School of Zhang Jian , Nantong University，Nantong , Jiangsu, 226019 China；

2. School of Transportation and Civil Engineering，Nantong University，Nantong , Jiangsu, 226019 China）

Abstract： To improve the accuracy and speed of human pose estimation, a channel-split-based human pose estimation algorithm, named Channel-Split Residual Steps Network (Channel-Split RSN), is proposed. First of all, Channel-Split Blocks are proposed to apply convolution operation for split feature in order to obtain rich feature representation. Then, Feature Enhancement Block are introduced to further split feature channel and employ different strategies for different groups which can reduce similar features in feature channels. Finally, to further enhance the Pose Refine Machine in Channel-Split RSN, combined with improved spatial attention mechanism, a Pose Refine Machine based on feature spatial correlation, named Context-PRM, is proposed. Experimental results show that on the COCO test-dev dataset, our algorithm reaches 75.9% AP and 55.36 FPS, and the Params(M) of the model is only 18.3. Compared with the traditional RSN18 and RSN50, the AP of the model is improved by 5% and 3.4%, respectively. FPS is 12.08 faster than the traditional RSN50. On the more challenging CrowdPose dataset, our approach achieves 66.9% AP and 19.16 FPS, an AP improvement of 4.6% compared to RSN18,which effectively improves the accuracy of human pose estimation and the model has a faster recognition speed. Our source code is available at https://github.com/qdd1234/Channel-Split-RSN.

Keywords： Channel-Split RSN；Human pose estimation；Channel-Split Block；Feature Enhancement Block；Context-PRM

基金项目：国家自然科学基金面上项目(61671255)；江苏省大学生创新训练计划项目(201910304158H)；江苏省大学生创新训练计划项目(202010304180H)；江苏省大学生创新训练计划项目(202010304122Y)。

作者简介：周昆阳(2000-)，男，江苏盐城人，主要研究方向为计算机视觉；赵梦婷(2001-)，女，江苏沭阳人，主要研究方向为图像处理；张海潮(2001-)，女，四川德阳人，主要研究方向为图像处理。邵叶秦(1978-)(通讯作者)，男，浙江海宁人，副教授，博士，主要研究方向为计算机视觉

收稿日期：2021-06-28　**修回日期：**　**E-mail：**hnsyk@ntu.edu.cn

0 引 言

人体姿态估计是人体运动识别、运动学分析、人机交互、动画制作等方面的基础性工作。人体姿态估计的目的是借助摄像头等传感器，在复杂场景、不同人群中对人体的关节点进行准确定位。

多年来，人体姿态估计大多基于手工特征，主要包括基于可穿戴设备[1]和基于模版匹配[2]的方法。这两种方法都存在泛化能力较低、检测流程繁琐等缺点。随着神经网络的发展，人体姿态估计取得很大的进展。Wei 等[3]提出了卷积姿态网络(Convolutional Pose Machine, CPM)，首次对人体关节点信息进行建模，基于模型输出的热力图，在每个通道上找到最大响应点实现人体姿态估计。同时，由于行人检测算法性能逐步提升，出现很多优秀的检测模型，例如 YOLO、Faster-RCNN 等。这使得姿态估计算法逐渐从单人姿态估计转向多人姿态估计。按照姿态估计实现方式的差异，多人姿态估计分为自底向上(Bottom-Up)和自顶向下(Top-Down)两类。

自底向上的多人姿态估计是先检测所有关节点，然后根据所属人体组装这些关节点，其典型模型是 OpenPose[4]。OpenPose 通过 CPM 定位所有关节点的位置，并采用部件亲和场(Part Affinity Fields)组装定位好的各个关节点。OpenPose 已经被许多学者广泛应用到各个领域中。Van-Hung Le 等[5]将 OpenPose 应用到传统武术表演评价中，唐心宇等[6]将 OpenPose 应用到患者康复医疗中。也有学者根据 OpenPose 的优缺点和应用场景对 OpenPose 进行改进。冯文宇等[7]提出 CT-OpenPose，通过将 OpenPose 的底层特征提取网络替换为带有软阈值的残差网络，并且改进了 OpenPose 底层特征提取的流程和模型压缩方式，在特征提取网络中加入权值修剪和跨层连接机制，有效提高了模型的检测速度和准确度，但 CT-OpenPose 仅在特定场景下对 OpenPose 模型进行了改进，通用性不高。

自顶向下的多人姿态估计是先通过人体检测器检测出人所在区域，然后在该区域上进行单人姿态估计。具有代表性的是旷世科技提出的级联金字塔网络(Cascaded pyramid network, CPN)[8]和 Fang 等提出的 RMPE[9]。级联金字塔网络是一种由粗到细的网络模型，通过利用单人的上下文信息完成人体姿态估计。RMPE 由对称空间变换网络(Symmetric Spatial Transformation Networks)，参数姿态非极大抑制(Non-maximum suppression)和姿态引导提议生成器(Pose Guide Proposal Generator)三个部分组成，通过处理不精确的人体定位框和检测冗余，有效地提高了人体姿态估计的准确率。这些主流人体姿态估计算法虽然提高了模型识别的准确率，但存在模型过大、预测速度较慢的问题，不利于实际使用。

本文提出一种基于通道切分的自顶向下的人体姿态估计算法 Channel-Split Residual Steps Network (Channel-Split RSN): **(1)**在传统残差阶梯网络(Residual Steps Network, RSN)基础上，提出

一种通道切分模块，将输入特征的通道分成 k 个部分，对 k-1 个部分分别使用卷积提取特征，再将 k 个特征沿通道进行拼接，并和通道切分模块的输入特征相融和，得到丰富的特征表示。**(2)** 在通道切分模块的基础上引入特征增强模块，使用分组卷积和逐点卷积，对特征通道不同部分采取不同的处理策略，以减少通道内的相似特征，提高特征提取的效果。**(3)**为了提高人体姿态估计的准确性，提出一种基于改进的空间注意力机制的姿态修正机 Context-PRM，通过考虑特征在空间上的相关性，提升姿态估计的准确率。实验结果表明，本文的方法能够有效地提高人体姿态估计的准确率，且模型具有较快的预测速度和较高的实用性。

1 残差阶梯网络算法介绍

残差阶梯网络[10]是旷世科技提出的人体姿态估计算法，算法采用自顶向下方法进行人体姿态估计，人体检测器使用 MegDetv2[11]。整体网络结构由多个残差阶梯网络级联而成，每个残差阶梯网络包含 8 个残差阶梯块(Residual Steps Block, RSB)，如图 1 所示。每个残差阶梯块通过密集的逐个元素相加方式加强特征的融合，有效地丰富了人体姿态估计的特征表示。

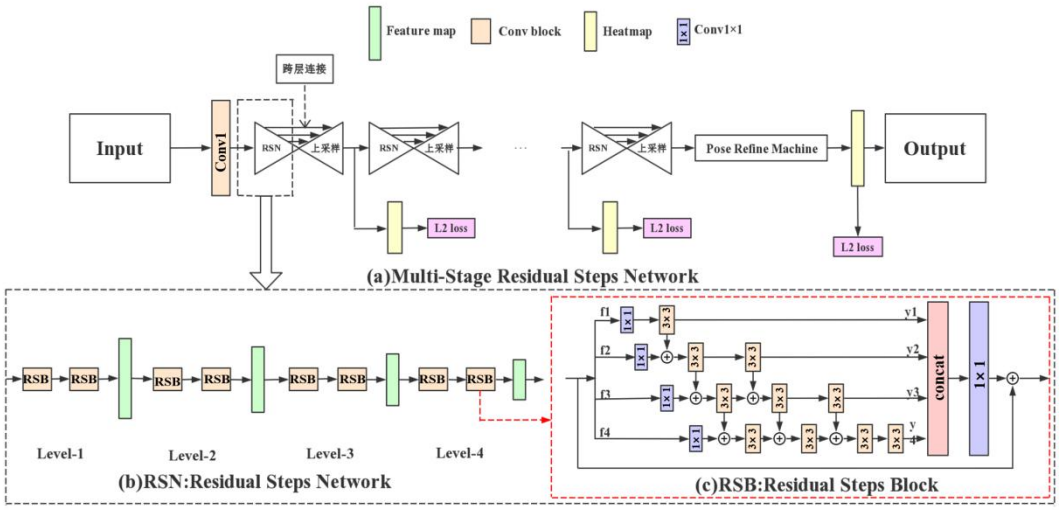


图 1 传统残差阶梯网络结构图

另外，残差阶梯网络还包含了一个基于注意力机制的姿态修正机(Pose Refine Machine, PRM)，如图 2 所示。输入特征经过 3×3 卷积后进入三条路径：第一条路径(First path)用于得到通道注意力，第三条路径(Third path)用于生成空间注意力，第二条路径(Second path)是将这两条路径生成的注意力作用在输入特征上。姿态修正机使用通道注意力和空间注意力进一步提高人体关节点的定位精度。

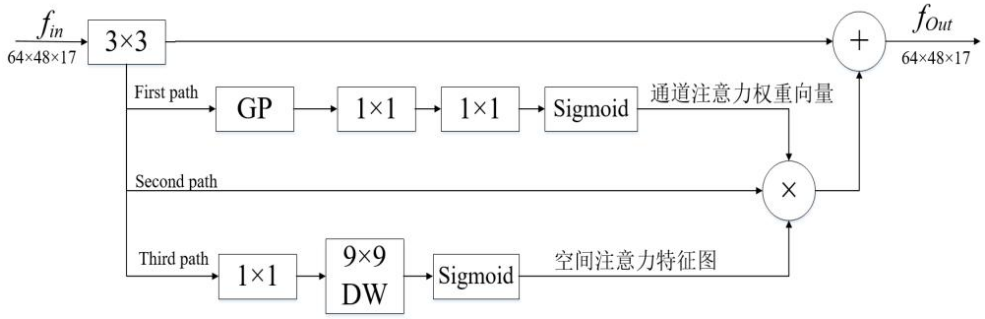


图 2 姿态修正机示意图，GP 代表全局平均池化，DW 代表深度可分离卷积

2 基于通道切分的人体姿态估计算法

2.1 算法思路

本文采用自顶向下的方式实现人体姿态估计，使用 RSN18 作为特征提取网络，YOLOv4[12] 作为人体检测器，算法结构如图 3 所示。输入图像首先经过卷积和最大池化提取特征，接着特征经过一个包含 8 个残差阶梯块的残差阶梯网络、跨层连接和上采样得到人体姿态估计的初步结果。需要注意的是，本文在最后一个残差阶梯块中引入通道切分模块，替换原有的 3×3 卷积，如图 3 中绿色方块所示。最后，基于改进的姿态修正机(Context-PRM)，得到更加准确的人体姿态估计。整体算法命名为 Channel-Split Residual Steps Network (Channel-Split RSN)。

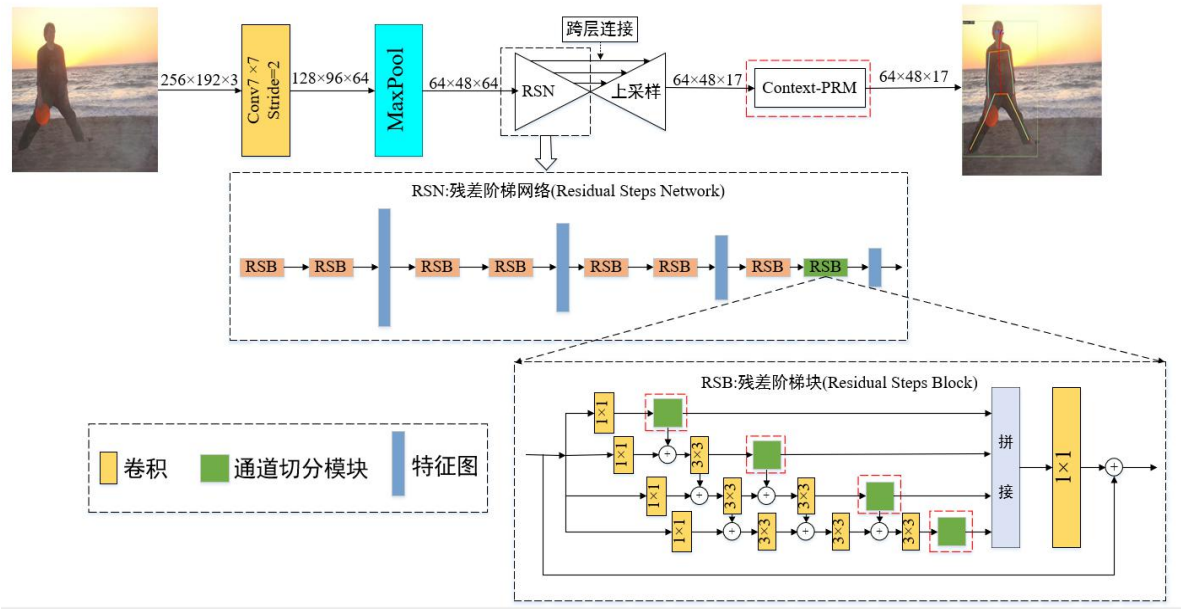


图 3 Channel-Split RSN 网络结构

2.2 通道切分模块

2.2.1 通道切分模块算法流程

为了得到丰富的特征表示，本文提出通道切分模块，将特征通道分成若干等份，分别提取每个部分的特征后，再融合起来(如图 4 所示)。

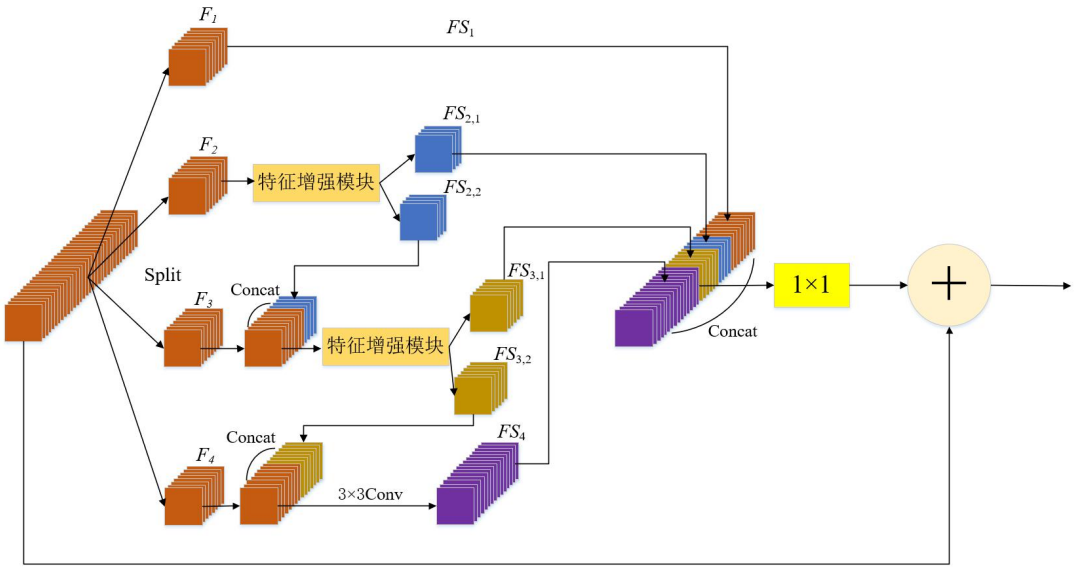


图 4 通道切分模块(k=4)，3×3Conv 代表 3×3 卷积+BN+ReLU

具体来说，输入特征首先沿通道平均分成 k 个部分，每个部分记为 $F_i(1\le i\le k,k>2)$。每一部分的特征进行如下操作：

$$FS_1 = F_1, i = 1 \tag{1}$$

$$FS_{2,1}, FS_{2,2} = H_2(F_2), i = 2 \tag{2}$$

$$FS_{i,1}, FS_{i,2} = H_i(F_i + F_{i-1,2}), 3 \le i \le k-1 \tag{3}$$

$$FS_k = Conv_{3\times3}(F_k + F_{k-1,2}), i = k \tag{4}$$

其中，$H_i(2\le i\le k-1)$ 表示特征增强模块，$Conv_{3\times3}$ 包括 3×3 卷积→批量归一化层(Batch Normalization, BN)→ReLU 激活函数，+号表示拼接(Concat)。

然后，将 FS_1、$FS_{i,1}(2\le i\le k-1)$、FS_k 沿通道拼接，经过一个 1×1 卷积后再和通道切分模块输入特征相加，得到整个通道切分模块的输出。这里，通道切分模块的输入和输出特征具有相等的尺寸和通道数。

通道切分模块通过切分特征通道并融合不同大小通道数的特征促进特征之间的信息交流，进而丰富通道切分模块中的特征表示。同时，通道切分模块使用特征增强模块减少特征通道内的相似特征，提高特征提取的效果。

2.2.2 特征增强模块

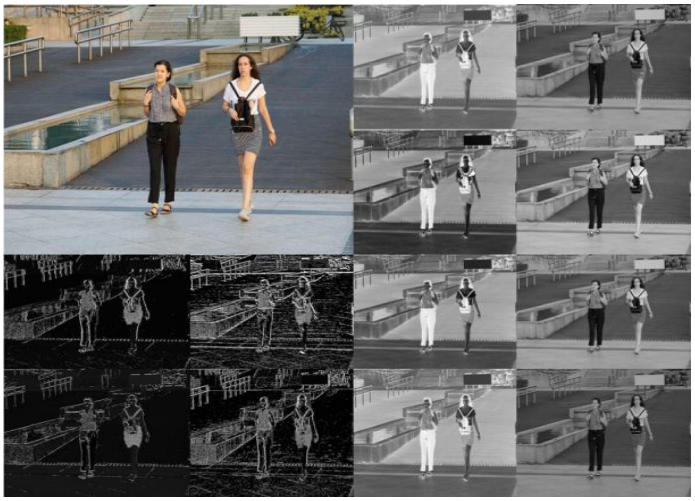
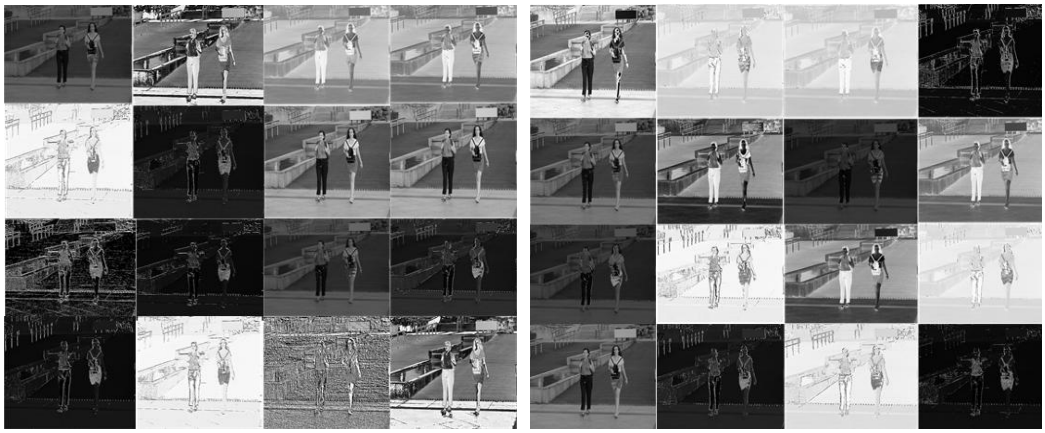


图 5 标准卷积的部分通道特征展示



(a) $y_{i,1}$ 通道特征图　　(b) $y_{i,2}$ 通道特征图

图 6 $y_{i,1}$ 和 $y_{i,2}$ 通道特征图

标准卷积对输入特征的所有通道进行相同的卷积操作。经过标准卷积后的特征中，部分通道的特征十分相似(如图 5 所示)，因此这些通道信息存在冗余，对相似特征进行特征提取会造成特征冗余且增加计算量。即使将特征通道(通道数为 32)简单的平均分成两个部分，如图 6 所示。得到的 $y_{i,1}$ 和 $y_{i,2}$ 这两个部分中仍有相似的特征，若对 $y_{i,1}$ 和 $y_{i,2}$ 使用相同处理方式，还会导致得到的特征通道中存在较多相似特征。

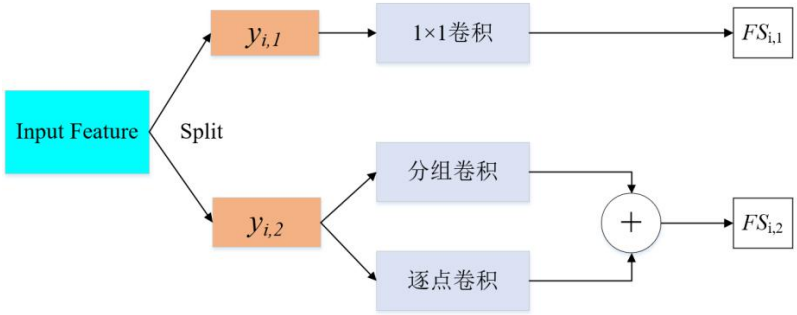


图 7 特征增强模块

为了减少通道内的相似特征，GhostNet[13]中提出对特征的每个通道进行对应的线性变换以此来减少通道内的相似特征。受到该思想的启发，本文设计一个特征增强模块，对不同的特征部分采用不同的处理策略，如图 7 所示。具体来说，将特征沿通道平均分成两个不同部分 $y_{i,1}$ 和 $y_{i,2}$，对 $y_{i,1}$ 使用 1×1 卷积提取特征以补充局部细节信息，而对 $y_{i,2}$ 使用 3×3 分组卷积和逐点卷积(pointwise convolution)[14]提取内在特征(如图 7 所示)。分组卷积将特征的通道分为 g 组(g>1)，每个组分别进行卷积操作。与标准卷积不同，分组卷积不同组的卷积核参数不同，相当于每个组进行不同的操作，因此每个组的输出特征是各不相同的，有效减少特征通道内的相似特征。但因为 $y_{i,2}$ 的通道信息中包含若干模块，每个模块都表示一个主要的特征(例如:条纹、颜色等)，分组卷积使得模块间的信息交流被隔断[15]。为此，本文增加逐点卷积，实现输入特征在通道方向上的加权组合，然后将分组卷积和逐点卷积提取的特征相加，丰富 $y_{i,2}$ 通道的特征表示。

2.2.3 通道切分模块计算复杂度分析

本文使用通道切分模块替代卷积核大小大于 2 的卷积操作。为了衡量通道切分模块的计算复杂度，本文对比 $s\times s(s>2)$标准卷积的计算复杂度(公式 5)和通道切分模块的计算复杂度(公式 9)，这里，标准卷积计算复杂度用 P_{normal} 表示，通道切分模块计算复杂度用 P 表示。

$$P_{normal}=s\times s\times k\times w\times k\times w \tag{5}$$

其中，k 为特征被通道切分模块分成的份数，$k>2$，w 为每个部分的通道数。

每个特征增强模块对应的计算复杂度 P_i 如下:

$$P_i=(s\times s\times\frac{1}{4g}+\frac{1}{2})\times w\times w\times(1+\frac{2^{i-2}-1}{2^{i-2}}) \tag{6}$$

其中，$2\le i\le k-1$，g 为分组卷积的分组数，$g>1$。当 $i=1$ 时，通道切分模块计算复杂度为 0，当 $i=k$ 时，对应的 3×3 卷积，其计算复杂度 P_k 如公式(7)所示:

$$P_k=3^2\times w^2\times(\frac{2^{k-2}-1}{2^{k-2}}+1)^2 \tag{7}$$

通道切分模块中 1×1 卷积的计算复杂度如公式(8)所示。

$$P_{1\times1} = w^2 \times k^2 \tag{8}$$

因此，整个通道切分模块的计算复杂度 P 如下：

$$P=(s\times s\times\frac{1}{4g}+\frac{1}{2})\times w\times w\times\sum_{i=2}^{k-1}(\frac{2^{i-2}-1}{2^{i-2}}+1)^2+3^2\times w^2\times(\frac{2^{k-2}-1}{2^{k-2}}+1)^2+w^2\times k^2 \tag{9}$$

$$\therefore P<(s^2\times\frac{1}{4g}+\frac{s^2}{2})\times w^2\times\sum_{i=2}^{k-1}(1+\frac{2^{i-2}-1}{2^{i-2}})^2+3^2\times w^2\times2^2+k^2\times w^2$$

$$=\succ\prec(\frac{s^2}{4}+\frac{s^2}{2})w^2\times\sum_{i=2}^{k-1}(1+\frac{2^{i-2}-1}{2^{i-2}})^2+s^2\times w^2\times2^2+k^2\times w^2$$

$$=\succ\prec s^2\times w^2\times\sum_{i=2}^{k-1}(1+1)^2+s^2\times w^2\times2^2+k^2\times w^2$$

$$=\succ\prec s^2\times w^2\times\sum_{n=1}^{k-2}(1+1)^2+s^2\times w^2\times2^2+k^2\times w^2$$

$$=\succ\prec w^2\times\left[4\times(k-2)\times s^2+s^2\times4+k^2\right]$$

$$=\succ\prec w^2\times\left[s^2\times(4k-4)+k^2\right]$$

$\because$ 当 k>2,s>2 时 $(k-2)\times s-k\geq0$

又 $\because k^2\times s^2-s^2\times(4k-4)-k^2$ 可以分解因式为 $[(k-2)\times s+k]\times[(k-2)\times s-k]$

$\therefore k^2\times s^2-s^2\times(4k-4)-k^2\geq0$ 成立

$\therefore w^2\times\left[s^2\times(4k-4)+k^2\right]\leq w^2\times k^2\times s^2$

$\because P<w^2\times\left[s^2\times(4k-4)+k^2\right]$

$\therefore P<s^2\times w^2\times k^2=P_{normal}$

因此，本文提出的通道切分模块相较于标准卷积计算复杂度更小，需要更少的计算资源。

2.3 Context-PRM 算法

残差阶梯网络使用一种基于注意力的姿态修正机，如图 2 所示。本文在姿态修正机的基础上，对姿态修正机的空间注意力机制进行改进，提出了 Context-PRM，如图 8 所示。

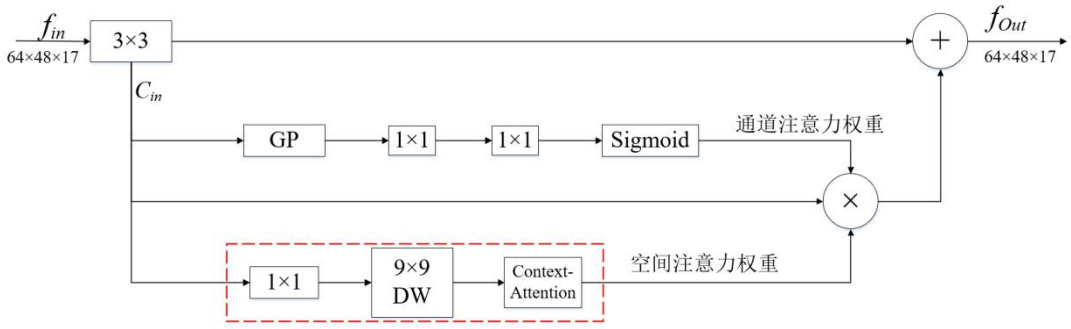


图 8 Context-PRM 结构，GP 表示全局平均池化，DW 表示深度可分离卷积

Context-PRM 的空间注意力机制如公式(10)所示：

$$attention_{space} = ContextAttention(DW(Conv_{1\times1}(C_{in}))) \tag{10}$$

其中，C_{in} 表示空间注意力机制的输入特征，$Conv_{1\times1}$ 表示 1×1 的卷积操作，DW 表示深度可分离卷积 ，$ContextAttention$ 为本文提出的 Context-Attention 模块。Context-Attention 模块的具体结构如图 9 所示。

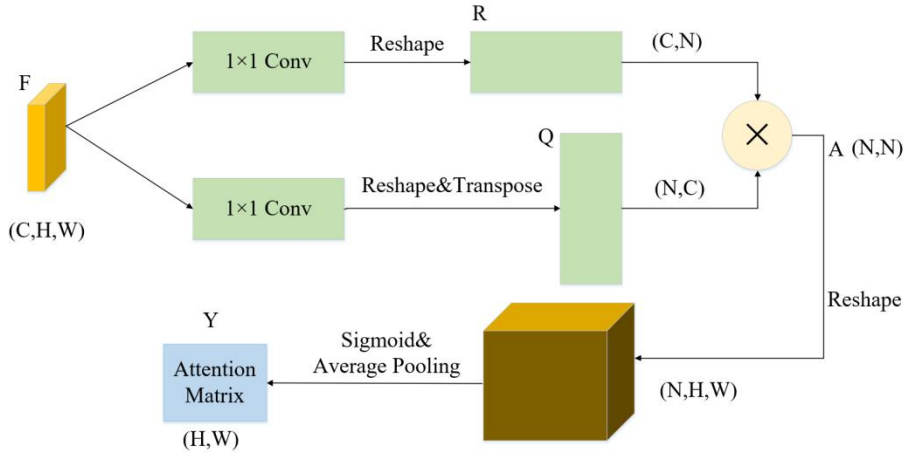


图 9 Context-Attention 结构

输入特征首先分别经过两个 1×1 卷积，之后将输出特征的尺寸重塑(reshape)成二维，得到两个输出 R、Q，其中 N=H×W。Q 进行转置，尺寸变成(N,C)。为了得到两个特征 R 和 Q 间的关联性，这里构造一个关联特征 A=Q×R。接着，将 A 重塑到 N×H×W 的三维矩阵。为了将关联特征 A 归一化，本文采用平均池化+Sigmoid 激活函数，将 A 的维度从三维变成二维，得到大小为 H×W 的注意力矩阵(Attention Matrix)。最后通过逐个元素相乘将注意力矩阵作用于第二条路径的特征。

3 实验结果与分析

3.1 实验数据与平台

3.1.1 实验数据

本文数据集使用 COCO 数据集[16]和 CrowdPose 数据集[17]。COCO 数据集的训练集 COCO train2017 包含 50K 张行人图片和 150K 个人体标注实例。验证集 COCO minival dataset 包含 5K 张图片，测试集 COCO test-dev 包含 20K 图片，模型输入图像大小为 256×192。

CrowdPose 数据集包含较多拥挤的场景，相较于 COCO 数据集更具挑战性。其中训练集包括 10K 张图片，验证集包括 2K 张图片，测试集包括 20K 张图片，模型输入图像大小为 512×512。

数据预处理：本文对数据集进行了数据增广，具体操作包括随机旋转、随机缩放、随机亮度调整、随机对比度调整、随机饱和度调整等图像增强方式。随机缩放的比例因子为 0.7~1.35；随机旋转角度为-45 度~+45 度；随机亮度调整首先设定阈值为 0.5，然后随机在区间(0,1)内抽取一个数 a，如果 $a\geq0.5$，则亮度调整比例为 a，如果 $a<0.5$，则在区间$(-a,a)$内随机抽一个数 b，调整比例即为 b+1。随机对比度调整、随机饱和度调整和随机亮度调整方法相同，阈值均取 0.5。

本文方法的超参数如下:通道切分模块 k=4、特征增强模块中分组卷积分组数 g=2、COCO 数据集共训练 200 轮，CrowdPose 数据集共训练 250 轮、单批次训练样本数量(batch_size)为 20。参数 k 值和分组卷积分组数 g 的确定将在实验部分进行详细阐述。本文使用基于 Object Keypoint Similarity (Oks)[18]的平均准确率(Average Precision, AP)作为模型准确度的评价标准，每秒预测图片数量(Frames Per Second, FPS)作为模型预测速度的评价标准，Params(M)作为模型大小的评价指标。Oks 定义如公式(11)所示。

$$\mathrm{Oks}=\frac{\sum_i \exp\{-d_{p^i}{}^2/(2S_p{}^2\times\sigma_i{}^2)\}\delta(v_{p^i}=1)}{\sum_i \delta(v_{p^i}=1)} \tag{11}$$

其中，p 表示某个人的 id，p^i 表示某个人的第 i 个关节点，d_{p^i} 表示预测的第 i 个关节点和真实关节点间的欧氏距离，v_{p^i} 表示这个关节点在图片上是否可见，S_p 表示这个人所占面积的平方根(根据人的标注框计算得到)，σ_i 表示第 i 个关节点的归一化因子，它是第 i 个关节点在数据集中坐标的标准差，σ_i 越大，说明这个关节点在数据集上的坐标变化大; 否则说明它的坐标变化小。对于 $\delta(v_{p^i}=1)$，如果 $v_{p^i}=1$ 成立，那么 $\delta(v_{p^i}=1)=1$，否则 $\delta(v_{p^i}=1)=0$，这里表示仅计算真实值中已标注的关节点。计算基于 Oks 的平均准确率的步骤为:先设定阈值 th，若某幅图片计算的 Oks 值大于 th，表明该图片关节点检测有效，否则无效。本文的 th=0.95。基于所有图片的检测结果，计算平均准确率 AP 如公式(12)所示。

$$\mathrm{AP}=\frac{\text{关节点检测有效的图片}}{\text{所有图片}} \tag{12}$$

FPS 定义如公式(13)所示。

$$\mathrm{FPS}=\frac{\text{预测图片总数}}{\text{预测总时间}} \tag{13}$$

Params(M)定义如下：

$$\mathrm{Params(M)}=\frac{\text{模型大小(MB)}}{4} \tag{14}$$

3.1.2 实验平台

模型训练与测试在百度 AI Studio 平台进行，CPU 是 Intel(R) Xeon(R) Gold 6271C @ 2.60GHz，GPU 为 Tesla V100 显存 16GB，内存 32GB。编程环境为 Python3.7，深度学习框架为 PaddlePaddle 1.8.4。

3.2 训练策略

本文在 4 个 Tesla V100 GPU 上进行训练，优化方法选择 Adam。为了加快模型的收敛，本文选择余弦学习率和指数移动平均的训练策略(Exponential Moving Average, EMA)。学习率和总训练轮数(*epochs*)的关系如公式(15)所示。

$$learning\ \_rate = begin\ \_rate \times 0.5 \times (\cos(\frac{epoch \times \pi}{epochs}) + 1) \tag{15}$$

其中，*begin_rate*=0.0005 为初始学习率，*epoch* 为当前训练轮数，*epochs* 为总的训练轮数。

为了使得模型参数平缓更新，本文在模型训练时采用指数移动平均策略。指数移动平均通过指数衰减方式计算参数更新过程中的移动平均值。对于每一个参数 W，都有一个指数移动平均值 W_t，如公式(16)。

$$\mathrm{W_t}=\alpha\times\mathrm{W_{t\text{-}1}}+(1-\alpha)\times\mathrm{W(t\geq 1)} \tag{16}$$

其中，α=0.998 为衰减系数，t 表示迭代次数，W_0 为 0。

此外，本文在训练模型时也采用迁移学习策略，先在公开数据集 ImageNet[19]上得到预训练权重，接着迁移到本文数据集上。预训练时，模型的超参数与训练时保持一致。

3.3 实验结果分析

3.3.1 通道切分模块和特征增强模块作用

为了验证通道切分模块和特征增强模块的作用，本文在 RSN18 基础上依次加上通道切分模块和包含不同分组数的特征增强模块。从表 1 可以看出，相比于 RSN18，增加了通道切分模块的 Channel-Split RSN18 的 FPS 降低了 4.3，但 AP 提升了 1.12%。在使用特征增强模块后，分组卷积代替原来的 3×3 卷积，其分组数 g 分别取 2、4、6，模型 Channel-Split RSN18(g=2)、Channel-Split

RSN18(g=4)和 Channel-Split RSN18(g=6)的 AP 和 FPS 均高于未使用特征增强模块的 Channel-Split RSN18。相较于 RSN18，虽然 Channel-Split RSN18(g=2)、Channel-Split RSN18(g=4)和 Channel-Split RSN18(g=6)的 FPS 有所降低，但 AP 分别提升 1.46%、1.8%、1.69%。因此，本文提出的通道切分模块和特征增强模块对于模型检测准确率的改进是有效的。

表 1 通道切分模块和特征增强模块的作用(未使用注意力机制)。

g 表示特征增强模块中分组卷积的分组数。

特征提取网络模型	是否使用特征增强模块	AP/%	FPS
RSN18	×	70.55	**61.9**
Channel-Split RSN18	×	71.67	57.6
Channel-Split RSN18(g=2)	√	72.01	57.7
Channel-Split RSN18(g=4)	√	**72.35**	58.1
Channel-Split RSN18(g=6)	√	72.24	58.6

此外，通过对比分组卷积的不同分组数（g=2、4、6）可以看出，使用 4 个分组的 Channel-Split RSN18(g=4)相较于使用 2 个分组的 Channel-Split RSN18(g=2)，其 AP 和 FPS 分别增加 0.34%、0.4，而相较于使用 6 个分组的 Channel-Split RSN18(g=6)，虽然其 FPS 降低了 0.5，但是 AP 提高 0.11%。在 FPS 差不多的情况下，考虑到算法在 AP 上的性能，本文选择 g=4 的特征增强模块。

3.3.2 Context-PRM 作用

为了衡量改进的空间注意力机制对姿态修正机的影响，本文实验对比了改进的姿态修正机(Context-PRM)和传统的姿态修正机(PRM)。

表 2 Context-PRM 的作用

姿态修正机模型	特征提取网络	AP/%	Params(M)	FPS
PRM	Channel-Split RSN18(g=4)	73.92	15.7	57.14
Context-PRM	Channel-Split RSN18(g=4)	**75.81**	17.3	55.36

从表 2 可以看出，将 Channel-Split RSN 的姿态修正机替换为 Context-PRM 后，虽然模型检测速度有所下降，但模型 AP 提高了 1.89%。

3.3.3 与主流注意力机制的对比

本文使用现阶段其他的主流注意力机制代替 Context-PRM 实现姿态修正机，并对比了它们的性能，如表 3 所示。

表 3 不同注意力机制对比

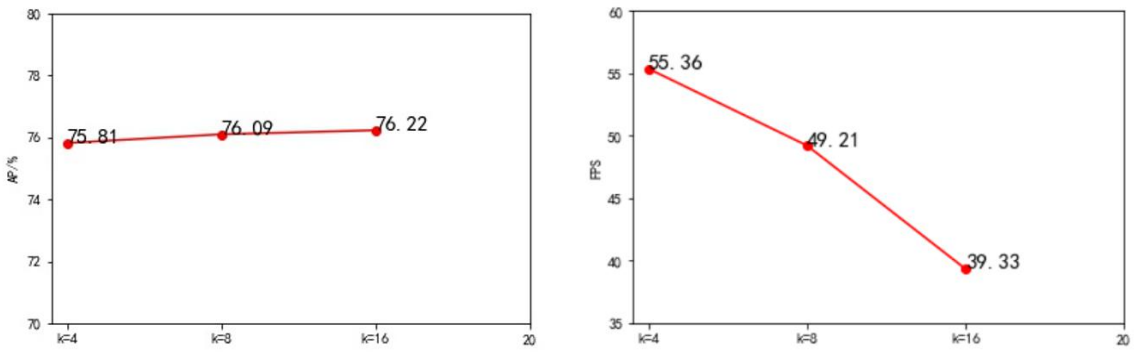
注意力机制	特征提取网络	AP/%
None	RSN18	70.55
CBAM	RSN18	69.83
SE-block	RSN18	70.45
PRM	RSN18	72.14
Context-PRM	RSN18	**72.29**
None	RSN50	74.32
PRM	RSN50	74.70
Context-PRM	RSN50	**74.88**

Convolutional Block Attention Module(CBAM)[20]和 SE-block[21]是具有代表性的注意力机制。从表 3 结果可以看出，在使用 CBAM 和 SE-block 后，RSN18 的 AP 分别下降 0.72%和 0.1%。而改进的姿态修正机(Context-PRM)和传统的姿态修正机(PRM)在模型的 AP 上明显优于 CBAM 和 SE-Block。

另外，本文也对比了不同的特征提取网络 RSN18 和 RSN50。结果表明，在 PRM 上，RSN18 和 RSN50 的 AP 分别提高 1.59%和 0.38%；在 Context-PRM 上，RSN18 和 RSN50 的 AP 分别提高 1.74%和 0.56%。实验结果表明，Context-PRM 比 PRM 对于模型 AP 提升更有效，同时相较于容量较大的 RSN50 模型，Context-PRM 对于容量较小模型 RSN18 的 AP 提升更明显。

3.3.4 通道切分模块的 *k* 值选择

不同的 *k* 值对模型的准确率和预测速度有不同的影响，本文通过控制变量法确定 *k* 值，如图 10 所示。本文输入通道切分模块的通道数为 512。通道切分模块将特征分成 *k* 个通道数相等的部分，即 $k=2^n$ (n>1，*k*>=4)。从图 10 中的实验结果可知，当 *k*=8 和 *k*=16 时，虽然 AP 比 *k*=4 提升 0.28%和 0.41%，但 FPS 比 *k*=4 分别低 6.15 和 16.03。在 AP 差不多的情况下，考虑到算法在 FPS 上的性能，本文选择 *k*=4。



(a) *k* 与准确率曲线　　(b) *k* 与 FPS 曲线

图 10 *k*、准确率、FPS 三者关系曲线

3.3.5 COCO test-dev 数据集上与主流姿态估计算法对比

前面的对比实验均在 COCO minival dataset 上进行。为了实验对比的公平性，本文将 Channel-Split RSN 和主流姿态估计算法在 COCO test-dev 上进行对比，实验结果见表 4

表 4 COCO test-dev 测试结果。
预训练表示在 ImageNet 上预先训练过模型，DA 表示使用本文的数据增广方法。

模型	是否 预训练	输入 图片大小	Params(M)/M	AP/%	FPS
Openpose	-	-	-	61.8	-
G-MRI[22]	-	353×257	42.6	64.9	-
G-RMI[23]	-	353×257	42.6	68.5	-
文献[24]	×	384×288	-	71.8	-
CPN	√	384×288	58.8	72.1	-
SimpleBase	√	384×288	68.6	73.7	-
文献[17]	×	256×192	-	74.2	-
HRNet-W32[25]	√	384×288	28.5	74.9	-
文献[26]	-	384×288	29.5	75.2	-
HRNet-W48	√	384×288	63.6	75.5	-
RSN18	×	256×192	**12.9**	70.9	**59.72**
RSN50	×	256×192	25.7	72.5	43.28
本文方法 w/o DA	×	256×192	18.3	74.9	55.36
本文方法	√	256×192	**18.3**	**75.9**	**55.36**

从表 4 的实验结果可以看出，与主流人体姿态估计算法相比，本文方法在平均准确率上超过现阶段主要姿态估计算法，并且模型的 Params(M)低于现阶段多数主流人体姿态估计算法。另外，即使未使用预训练和数据增广策略，本文方法的 AP 达到 74.9%，仅比 HRNet-W48 低 0.6%，比 CPN 提高 2.8%，比 RSN18 提高了 4%，比 RSN50 提高了 2.4%，而模型的 Params(M)为 18.3M，比 HRNet-W48 低 45.3M，比 CPN 低 40.5M，比 RSN50 低 7.4%，FPS 提升了 12.08。使用数据增广和预训练后，本文方法的 AP 达到 75.9%，比 HRNet-W48 高 0.4%，比 RSN18 高 5%，比 RSN50 高 3.4%。以上实验结果表明本文提出的方法是有效的。

3.3.6 CrowdPose 数据集上与主流姿态估计算法对比

为了验证本文算法具有较高的鲁棒性和泛化能力，本文将 Channel-Split RSN 与主流人体姿态估计算法在更具挑战性的 CrowdPose 数据集上进行对比，实验结果如表 5 所示。其中，CrowdPose 数据集上本文方法的超参数设置与 COCO 数据集一致。

表 5 CrowdPose test-dev 实验结果，DA 表示使用本文的数据增广方法。

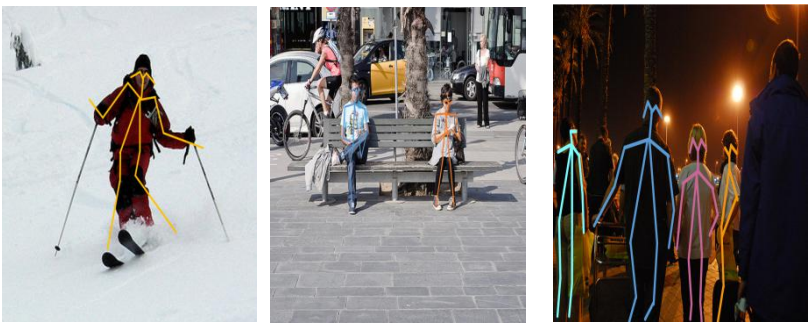
模型	是否 预训练	输入图片大小	Params(M)/M	AP/%	FPS
Mask-RCNN[27]	-	512×512	-	57.2	-
HrHRNet[28]	-	512×512	-	65.9	-
DEKR[29]	-	512×512	-	65.7	-
RSN18	×	512×512	**16.3**	62.3	**29.78**
RSN50	×	512×512	29.2	64.9	10.1
本文方法 w/o DA	×	512×512	27.6	66.1	19.16
本文方法	√	512×512	**27.6**	**66.9**	**19.16**

从表 5 实验结果可以看出，在 CrowdPose 数据集上本文方法在平均准确率上超过主流人体姿态估计算法。未使用预训练和实验数据增广策略，本文方法达到 66.1%的平均准确率，比 Mask-RCNN 高 8.9%，比 HrHRNet 提高 0.2%，比 RSN18 提高 3.8%。使用数据增广的策略后，本文方法的平均准确率达到 66.9%，比 DEKR 提高 1.2%，比 RSN50 提高 2%；FPS 达到 19.16，比 RSN50 高 9.06。以上实验结果均表明，本文方法在复杂的 CrowdPose 数据集上性能优于主流人体姿态估计算法。

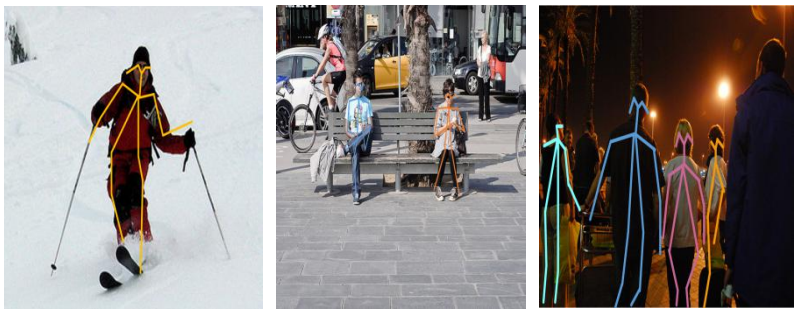
3.4 实验结果可视化

本文分别将 COCO 数据集和 CrowdPose 数据集上的检测结果进行可视化。在 COCO test-dev 数据集上，本文可视化了本文方法、CPN、SimpleBase、HRNet-W32 四类人体姿态估计算法的实验结果。本文用直线将检测到的人体各个相邻关节点连接起来，不同颜色的线表示不同的人，如图 11 所示。实验结果显示，在单人场景下，本文方法的关节点定位比 CPN 和 SimpleBase 更加准确。多人场景下，相较于 CPN、SimpleBase、HRNet-W32，本文方法能够对更多的人进行姿态估计。此外，在光线较暗的情况下，本文方法对多人人体姿态估计结果优于 CPN、SimpleBase 和 HRNet-W32。可视化结果表明，本文方法在人体姿态估计方面准确率高、鲁棒性强。

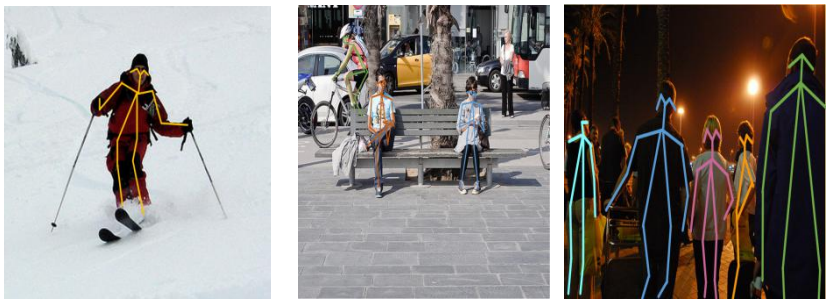
单人场景　　多人场景　　光线较暗场景



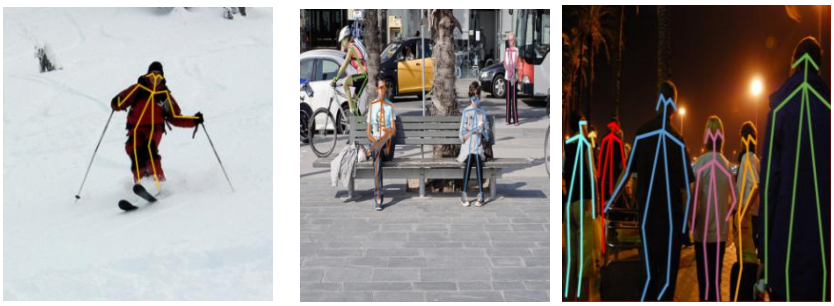
图(a) CPN 结果图



图(b) SimpleBase 结果图



图(c) HRNet-W32 结果图



图(d) 本文方法结果图

图 11 COCO 数据集姿态估计结果图

在 CrowdPose 数据集上，本文可视化了本文方法、Mask-RCNN、HrHRNet 三类人体姿态估计方法的实验结果，如图 12 结果所示。CrowdPose 数据集中主要以拥挤场景为主，本文从 CrowdPose 数据集中选择人在图像边缘(边缘场景)、人在图像中较远距离(目标较远场景)、人与人相互遮挡(遮挡场景)三类典型场景进行可视化。从边缘场景的可视化结果可以看出，相较于 Mask-RCNN 和 HrHRNet，本文方法能够有效地对图像边缘的人进行姿态估计。从目标较远的场景中可以看出，本文方法能够有效对图像中较远的人进行姿态估计。在遮挡场景中，本文方法相较于 Mask-RCNN 和 HrHRNet，能够有效对被遮挡的人进行姿态估计。以上实验结果表明本法方法在多人的复杂场景下仍然具有较好的鲁棒性和较高的检测准确率。

边缘场景　　目标较远场景　　遮挡场景

图(a) Mask-RCNN 结果图

图(b) HrHRNet 结果图

图(c) 本文方法结果图

图 12 CrowdPose 数据集姿态估计结果图。红色方框代表未能实现检测。

4 结语

4.1 总结

本文基于通道切分模块提出一种改进的人体姿态估计模型 Channel-Split RSN。首先通过通道切分模块增强卷积特征提取能力，同时减少卷积的计算复杂度；接着通过特征增强模块减少特征通道的相似特征以获得更加丰富的特征表示；最后提出一种改进的姿态修正机 Context-PRM，用于获得更准确的人体姿态关节点检测结果。在 COCO test-dev 上的实验表明，本文方法的 AP 达到 75.9%，FPS 达到 55.36，相较于 RSN18，AP 提高了 5%。与现阶段主流姿态估计算法相比，本文方法在 AP 上优于主流姿态估计算法，且模型的 Params(M)低于多数主流姿态估计算法。在更具挑战性的 CrowdPose 数据集上，本文方法达到 66.9%的 AP，模型性能优于主流人体姿态估计算法。

4.2 展望

人体姿态估计是一个具有挑战性的问题，本文提出的 Channel-Split RSN 在 COCO 数据集和 CrowdPose 数据集上达到不错的效果。今后将从网络结构、注意力机制上进一步改进本文方法，使其具有更好的性能，并可应用到更多实际场景。

参考文献

[1] LIU SH Q, ZHANG J CH, ZHU R. A wearable human motion tracking device using micro flow se nsor incorporating with micro accelerometer [J].IEEE Transactions on Biomedical Engineering, 2020, 67(4): 940-948.

[2] 王恬,李庆武,刘艳,等.利用姿势估计实现人体异常行为识别[J]. 仪器仪表学报, 2016,037(10): 2366-237 2.

[3] WEI S H, RAMAKRISHNA V, KANADE T, et al. Convolutional Pose Machines[C].IEEE Conferenc e on Computer Vision and Pattern Recognition (CVPR), 2016: 4724-4732.

[4] CAO Z, SIMON T, WEI S H, et al. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields [C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 7291-7299.

[5] VAN-HUNG L,TUONG-THANH N,NGOC-ANH T, THANH-CONG P. OpenPose's Evaluation in The Video Traditional Martial Arts Presentation[C].International Symposium on Communications and Information Technologies(ISCIT),2019: 76-81

[6] 唐心宇,宋爱国.人体姿态估计及在康复训练情景交互中的应用[J].仪器仪表学报,2018,39(11): 195-203.

[7] 冯文宇,朱洪堃,殷佳炜,等.无人 CT 智能姿态识别算法研究[J]. 仪器仪表学报, 2020, 41(08): 188-195.

[8] CHEN, YILUN, et al. Cascaded pyramid network for multi-person pose estimation[C]. Proceedings o f the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018: 7103-7112.

[9] FANG H S, XIE S Q, TAI Y W, et al. RMPE: Regional Multi-Person Pose Estimation[C].Internatio nal Conference on Computer Vision, 2017: 2353-2362.

[10] CAI Y H, WANG Z C, LUO Z X, et al. Learning Delicate Local Representations for Multi-Person Pose Estimation[C].European Conference on Computer Vision, 2020: 455-472.

[11] LI Z M, MA Y C, CHEN Y K, et al. Joint COCO and Mapillary Workshop at ICCV 2019: COC O Instance split Challenge Track[EB/OL](2020-10-06)[2020-12-26].https://arxiv.org/abs/2010.02475.

[12] ALEXEY B,CHIEN-YAO W, HONG-YUAN Mark L:YOLOv4:Optimal Speed and Accuracy of Objec t Detection[EB/OL](2020-04-23)[2020-12-28].https://arxiv.org/abs/2004.10934.

[13] HE K, YUNHE W, QI T, JIANYUAN G, CHUNJING X, CHANG X. GhostNet: More Features From Cheap Operations[C].The IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2020:1577-1586.

[14] BINH SON H, MINH-KHOI T, SAI-KIT Y. Pointwise Convolutional Neural Networks[EB/OL].(2021 -03-29)[2017-12-14].https://arxiv.org/abs/1712.05245.

[15] ZHANG T, QI G J, XIAO B, et al. Interleaved group convolutions[C].International Conference on Computer Vision(ICCV), 2017: 4383-4392.

[16] LIN T, MAIRE M, BELONGIE, S J., HAYS J, PERONA P, RAMANAN D, DOLLR P, ZITNICK C L: Microsoft coco: common objects in context[C].European Conference on Computer Vision(ECCV), 2014: 740-755.

[17] Li J F, Wang C, ZHU H, et al. CrowdPose: Efficient Crowded Scenes Pose Estimation and A New Benchmark[EB/OL].(2021-07-10)[2019-01-23]. https://arxiv.org/abs/1812.00324.

[18] 王柳程,欧阳城添,梁文.基于改进特征金字塔网络的人体姿态跟踪[J/OL].计算机工程:1-9[2021-05-08].https://doi.org/10.19678/j.issn.1000-3428.0058544.

[19] DENG J, DONG W, R.Socher, et al. ImageNet: A Large-Scale Hierarchical Image Database[C].The I EEE Conference on Computer Vision and Pattern Recognition(CVPR), 2009: 248-255.

[20] SANGHYUN W, JONGCHAN P, JOON-YOUNG L, et al. CBAM: Convolutional Block Attention M odule[C].European Conference on Computer Vision(ECCV), 2018:3-19.

[21] HU J, SHEN L, ALBANIE S, et al. Squeeze-and -excitation networks[J]. IEEE Transactions on Patt ern Analysis and Machine Intelligence. 2020,42(8):2011-2023.

[22] PAPANDREOU G, ZHU T, KANAZAWA N, et al. Towards accurate multipers on pose estimation in the wild[C].The IEEE Conference on Computer Vision and Pattern Recognition, 2017: 3711-3719.

[23] 胡保林.基于深度学习的人体关节点检测[D].电子科技大学信号与信息处理,2019.

[24] XIAO B, WU H, WEI Y. Simple baselines for human pose estimation and tracking[C].The European Conference on Computer Vision(ECCV), 2018:472-487.

[25] SUN K, XIAO B, LIU D, WANG J. Deep high-resolution representation learning for human pose estimation[C].The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019:5693 -5703.

[26] 罗梦诗,徐杨,叶星鑫.融入双注意力的高分辨率网络人体姿态估计[J/OL].计算机工程:1-10[2021-05-08].https://doi.org/10.19678/j.issn.1000-3428.0060493.

[27] HE K, GKIOXARI G, DOLLAR P, et al. Mask r-cnn(C].Proceedings of the IEEE international confe rence on computer vision(CVPR),2017:.2961-2969.

[28] CHENG B, XIAO B, WANG J D, et al. HigherHRNet: Scale-Aware Representation Learning for Bottom-Up Human Pose Estimation[EB/OL].(2021-07-10)[2020-03-12].https://arxiv.org/abs/1908.10357.

[29] GENG Z G, SUN K, XIAO B, et al. Bottom-Up Human Pose Estimation Via Disentangled Keypoint

Regression[EB/OL].(2021-07-10)[2021-04-06]. https://arxiv.org/abs/2104.02300.