基于改进轻量级沙漏模型的 2D 单人姿态估计研究与应用

黄 晨[1] 童维勤[1] 戴 伟[1] 陈一民[2*] 邹一波[3] 翁康年[4] 吴志华[5]

[1](上海大学计算机工程与科学学院 上海 200444)

[2](上海建桥学院信息技术学院 上海 201306)

[3](上海海洋大学信息学院 上海 201306)

[4](上海交响乐团 上海 200031)

[5](上海大剧院管理有限公司 上海 200031)

摘 要 提出一种基于改进轻量级沙漏模型的 2D 单人人体姿态估计方法。使用逆残差卷积来构建改进的轻量级沙漏模型,从而降低参数数量与计算量,使用多尺度特征融合以提高轻量级模型在遮挡情况下的关键点检测能力。引入知识蒸馏方法,使得改进的模型在略微降低检测准确度时,能大幅降低训练和部署所需要的计算资源。MPII 数据集和实际应用中的检测结果表明,改进的轻量级沙漏模型能有效检测人体骨骼关键点,实时性好、鲁棒性强,能在一定程度上克服遮挡问题。

关键词 姿态估计 沙漏模型 轻量级模型 知识蒸馏

中图分类号 TP391 **文献标志码** A DOI:10.3969/j.issn.1000-386x.2024.10.029

RESEARCH AND APPLICATION OF 2D SINGLE-PERSON POSE ESTIMATION BASED ON IMPROVED LIGHTWEIGHT HOURGLASS MODEL

Huang Chen[1] Tong Weiqin[1] Dai Wei[1] Chen Yimin[2*] Zou Yibo[3] Weng Kangnian[4] Wu Zhihua[5]

[1](*School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China*)

[2](*College of Information Technology, Shanghai Jian Qiao University, Shanghai 201306, China*)

[3](*College of Information Technology, Shanghai Ocean University, Shanghai 201306, China*)

[4](*Shanghai Symphony Orchestra, Shanghai 200031, China*)

[5](*Shanghai Grand Theater Management Co., Ltd., Shanghai 200031, China*)

Abstract A 2D single-person pose estimation method based on improved lightweight hourglass model is proposed. The inverse residual convolution was used to construct an improved lightweight hourglass model, which effectively reduced the number of parameters and the amount of calculation. Multi-scale feature fusion was used to improve the key point detection ability of the lightweight model under occlusion. The introduction of the knowledge distillation method enabled the improved model to significantly reduce the computing resources required for training and deployment when the detection accuracy was slightly reduced. The detection results of the MPII data set and practical application show that the improved lightweight hourglass model can effectively detect the key points of human bones, with good real-time performance and strong robustness, and can overcome the occlusion problem to a certain extent.

Keywords Human pose estimation Hourglass model Lightweight model Knowledge distillation

---

收稿日期:2021 - 06 - 03。上海市科技创新行动计划项目(19511104502,18511105300,16511101200)。**黄晨**,博士生,主研领域:增强现实,虚拟现实,人机交互。**童维勤**,教授。**戴伟**,硕士生。**陈一民**,教授。**邹一波**,讲师。**翁康年**,博士。**吴志华**,高工。

0 引 言

人体姿态估计(Human Pose Estimation)是计算机视觉领域的研究热点,是行为识别、人机交互、动作捕捉等问题的重要研究基础,可应用于 AR/VR、舞台演出、运动分析等领域,具备重要的研究价值和广泛的应用前景。但现有的人体姿态估计领域的研究工作主要聚焦于提升人体骨骼关键点的检测精确度,且基于传统方法的人体姿态估计依赖于人工特征和局部检测器,易受遮挡、拍摄角度、光照等因素干扰[1]。自 AlexNet 赢得 ImageNet 挑战以来,卷积神经网络在计算机视觉领域得到了广泛应用,显著提升了人体骨骼关键节点的检测效果。目前,基于深度学习方法的人体姿态估计已成为主流[2]。堆叠沙漏网络[3]在多个尺度上提取特征并进行融合,使用中间监督模块提升网络性能,能有效获取人体骨骼关键点的空间关系信息。Sun 等[4]将高低分辨率网络与多个低分辨率网络并联,进行了多次多尺度融合,从而获得了信息更为丰富的高分辨率表示,有效提升了人体关键点检测效果。Ke 等[5]提出了一种用于人体姿态估计的具有鲁棒性的多尺度感知神经网络,通过组合跨尺度的特征热力图来匹配人体关键点的上下文特征,利用多尺度回归网络来全局优化多尺度特征的结构匹配,并引入结构感知损失函数来改善关键点与各邻接点之间的匹配问题。上述这些方法在不同尺度或不同分辨率上提取特征,可以获得丰富的局部和全局姿态信息,能有效提高姿态估计的准确率,但也加深了网络的深度,提高了模型的复杂度。

虽然基于卷积神经网络的方法显著提高了关键点的检测精度,但伴随模型的结构越来越复杂,需要的算力也日益增长,这限制了人体姿态估计在强调实时性或只具备有限计算资源场景中的应用。针对该问题,MobileNet[6-7]采用深度可分离卷积替代标准卷积操作,不仅减少了参数数量还兼顾了实时性和准确率,但该模型提取人体特征能力有限,在肢体遮挡等情况下无法有效提取人体骨骼关键点。Openpose 网络模型[8]通过自下而上方法进行多人姿态估计,使用非参数表示对关键点位置和方向进行编码,并学习人体各部位之间的空间约束关系,实现高精度的人体姿态估计,但算法复杂度高,计算资源消耗大。Luo 等[9]提出了一种端到端的人体姿态估计网络,引入了特征金字塔网络结构,并结合基于注意力机制的监测模块,实现了实时检测的效果,但对不可见关键点的检测效果比较差。

本文针对 2D 单人人体姿态估计问题,提出了一种基于改进轻量级沙漏模型的人体姿态估计方法。同时通过引入知识蒸馏方法[10],显著提高了关键点检测能力。实验结果表明,本文所提出的改进轻量级沙漏模型在有效降低模型参数数量和训练成本的基础上,能取得较好的关键点检测效果,实时性好、鲁棒性强。

1 基于知识蒸馏的轻量级沙漏模型

1.1 网络整体结构

本文通过引入逆残差卷积对标准沙漏模型进行轻量化改进,并使用知识蒸馏方法来提高轻量级沙漏模型网络的检测性能。整体网络结构如图 1 所示,本网络以 RGB 图像作为输入,在教师网络中,由卷积模块提取的特征经过堆叠的标准沙漏模块(Stack = 8)和监督模块,最终经过预测模块得到预测结果;学生网络与教师网络结构相似,但学生网络中使用轻量级沙漏模型模块(Lightweight Hourglass Module,LHM)代替标准沙漏模型模块,并且学生网络堆叠的沙漏模块数量较少(Stack = 4)。

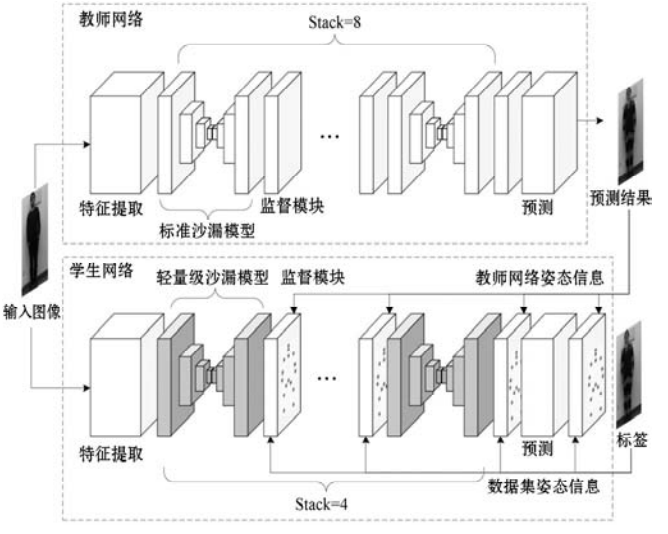


图 1 基于知识蒸馏的轻量级沙漏模型

使用同一网络模块进行堆叠串联的网络结构便于整体网络规模的伸缩控制。前一个 LHM 模块输出的带有骨骼关键点热力图的特征可以作为下一个 LHM 模块的输入,因而后续 LHM 模块可以充分利用已学习到的骨骼关键点之间的相对位置等信息,从而提高后续模块的预测精度。此外,有别于传统网络模型的 loss 计算方法,本网络设有中间监督模块,可以对每一个 LHM 模块进行预测。因此,本网络的 loss 取决于所有子模块的 loss。

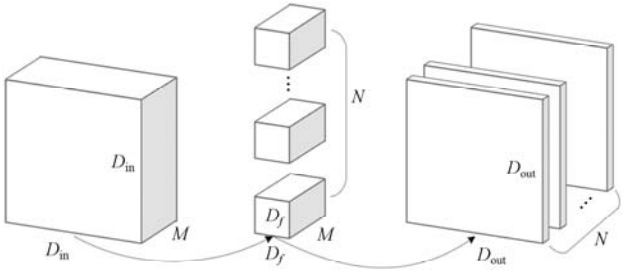
在使用知识蒸馏方法训练学生网络的过程中,预训练好的教师网络的预测结果会作为辅助标签和真实

值一起训练学生网络,这样不仅能加快学生网络的收敛,而且能提高学生网络对骨骼关键点的检测能力。

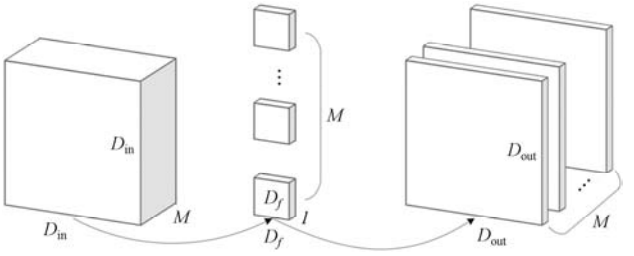
1.2　深度可分离卷积

如图 2(a)所示,对于标准卷积(Standard Convolution,SC),假设输入特征的尺寸为 $D_{in} \times D_{in} \times M$,卷积核尺寸为 $D_f \times D_f$,卷积核数量为 N,输出特征的尺寸为 $D_{out} \times D_{out} \times M$,则标准卷积的计算量可表示为:

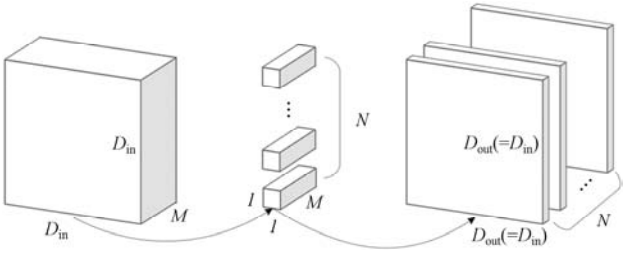
$$cost_{SC} = D_f \times D_f \times M \times N \times D_{in} \times D_{in} = D_f^2 \times D_{in}^2 \times M \times N \tag{1}$$



(a) 标准卷积



(b) 深度卷积



(c) 逐点卷积

图 2　不同类型的卷积操作

深度可分离卷积(Depthwise Separable Convolution,DSC)已在多个轻量化模型中得到了应用。如图 2(b)和图 2(c)所示,深度可分离卷积将标准卷积拆分为深度卷积(Depthwise Convolution,DC)和 1×1 逐点卷积(Pointwise Convolution,PC),针对移动终端等计算资源受限制场景使用 ReLU6 作为非线性激活函数,其计算公式为:

$$f(x) = \min(\max(x,0),6) \tag{2}$$

式中:x 为输入特征。

深度可分离卷积的计算量可表示为:

$$cost_{DC} = D_f \times D_f \times D_{in} \times D_{in} \times M,$$

$$cost_{PC} = 1 \times 1 \times M \times N \times D_{in} \times D_{in}$$

$$cost_{DSC} = cost_{DC} + cost_{PC} = D_f^2 \times D_{in}^2 \times M + D_{in}^2 \times M \times N \tag{3}$$

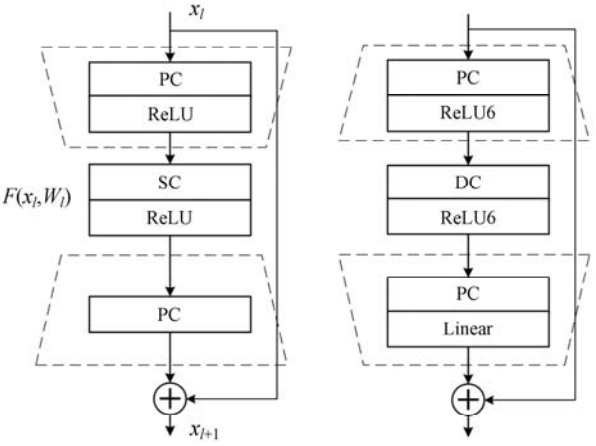
深度可分离卷积计算量与标准卷积计算量的比值可表示为:

$$p = \frac{cost_{DSC}}{cost_{SC}} = \frac{1}{N} + \frac{1}{D_f^2} \tag{4}$$

通常,卷积核数量 N 远大于卷积核尺寸 D_f,所以当卷积核为 3×3 时,深度可分离卷积的计算量约为标准卷积的 1/9。

1.3　逆残差模块

为在高维度特征中获得更丰富的信息,可以使用 PC 提高输入特征维度,然后使用 DC 提取特征,再使用 PC 进行降维,使得输入特征的维度与输出特征的维度保持一致。由于对低纬度特征进行卷积操作时,非线性激活函数 ReLU 会造成信息损失,因此在进行降维 PC 操作后使用线性激活函数代替 ReLU。同时,为了抑制多个模块堆叠造成的梯度消失问题,本文采用了与标准残差模块(Residual Block)类似的旁路设计。由于与标准残差模块中对特征维度先降后升的操作步骤相反,因此称其为逆残差模块(Inverted Residual Block),其结构如图 3 所示。



(a) 残差模块　　(b) 逆残差模块

图 3　残差模块与逆残差模块结构

残差模块可表示为:

$$x_{l+1} = x_l + F(x_l, W_l) \tag{5}$$

式中:$F(x_l, W_l)$ 为残差部分,包含两个 1×1 PC 卷积层和一个 SC 卷积层。

残差模块在第一层 PC 操作中将输入特征降低 n 倍,为保持输入特征与输出特征的维度不变,又在第三层 PC 操作中将特征维度提高 n 倍;类似的,逆残差模块在第一层 PC 操作中将输入特征提高 t 倍,在第三层 PC 操作中将特征维度降低 t 倍。假设两种模块都是对 128 维的特征进行处理,且 $n=2$,$t=2$,stride = 1,基本操作如表 1 所示。

表 1 残差模块与逆残差模块基本操作

模块	输入	操作	输出
残差模块	$D\times D\times 128$	1×1 PC, ReLU	$D\times D\times 64$
	$D\times D\times 64$	3×3 SC, ReLU	$D\times D\times 64$
	$D\times D\times 64$	1×1 PC	$D\times D\times 128$
逆残差模块	$D\times D\times 128$	1×1 PC, ReLU6	$D\times D\times 256$
	$D\times D\times 256$	3×3 DC, ReLU6	$D\times D\times 256$
	$D\times D\times 256$	1×1 PC, Linear	$D\times D\times 128$

根据式(1)、式(3),可得到逆残差模块和标准残差模块计算量的比值为:

$$p'=\frac{128\times 64\times D^2+3^2\times D^2\times 64\times 64+64\times 128\times D^2}{128\times 256\times D^2+3^2\times D^2\times 256+256\times 128\times D^2}\approx 0.8 \quad (6)$$

这表明逆残差模块能降低计算量,是实现网络轻量化的有效途径。并且逆残差模块在更高维度进行 DC 操作,能获得更丰富的空间信息。

1.4 轻量级沙漏模型

堆叠沙漏网络通过串联多个沙漏模型,利用多尺度特征实现了较好的姿态估计效果。但沙漏模型使用了大量的残差模块,需要的计算量较大。本文提出使用逆残差卷积模块代替沙漏模型中的标准残差模块来提高网络的特征提取性能。

参照沙漏模型,本文提出使用逆残差卷积模块构建了可以提取并融合多尺度特征的网络模块,如图 4 所示,本文使用的 4 阶轻量级沙漏模型可分为上采样和下采样两个过程,呈对称分布,并且模型的输入特征和输出特征的尺寸与通道数保持一致。

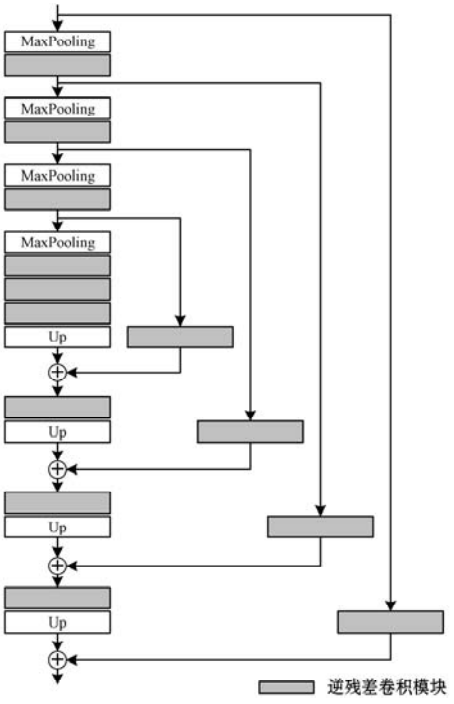


图 4 轻量级沙漏模型

在下采样阶段,网络分为两路,其中一路保留了原始特征的信息,另一路通过最大池化操作缩小输入特征尺寸,并利用逆残差模块提取特征。

在上采样阶段,为了将相邻且分辨率不同的信息能够融合在一起,本文使用最近邻抽样对低分辨率进行上采样,然后把相邻的两种尺度的特征相加生成融合特征。最终,模型经过 4 次上采样,得到了和原始输入特征尺寸相同的特征,实现多尺度信息的融合。本文提出的轻量级沙漏模型可以融合不同尺度的特征信息,能充分利用骨骼关键点的局部信息和各个关键点之间的关联信息,有较强的鲁棒性。

1.5 知识蒸馏

在知识蒸馏过程中,蒸馏函数和教师网络的教导函数是影响学生网络模型学习效果的重要因素。本文用了最小均方误差(Minimize Square Error, MSE)函数来评估教师网络和学生网络之间的差异性:

$$L_{ts}=\frac{1}{K}\sum_{k=1}^{K}\left\|m_k^s-m_k^t\right\|_2^2 \quad (7)$$

式中:m_k^t 表示经预训练后的教师网络模型对于第 k 个关键点的预测得到的置信图;m_k^s 表示训练过程中的学生网络对于第 k 个关键点的预测得到的置信图,K 为人体骨骼中的关键点总数。

本文认为学生网络在训练的不同阶段对来自教师网络和数据集知识的依赖度应有所区别。模型训练早期,学生网络的检测能力较弱,更为依赖教师网络的知识,从而加快网络的收敛;模型训练后期,学生网络已具备一定的检测能力,更侧重从数据集上学习知识。因此本文提出一种随迭代次数调整教师网络教导能力的教导函数:

$$L=\left(0.6-\sigma\times\frac{2i}{I}\right)L_{ts}+\left(0.4+\sigma\times\frac{2i}{I}\right)L_{mse} \quad (8)$$

式中:L_{ts}表示学生网络通过教师网络学习得到的知识;L_{mse}表示学生网络通过数据集学习得到的知识;σ 表示衰减系数,取值为 0.1;i 表示迭代次数。

2 实验方法

2.1 数据集介绍

本文使用 MPII 数据集对模型进行训练和测试。MPII 数据集是进行人体姿态估计的主流基准数据集。该数据集从 YouTube 视频中提取图像,包含大约 2.5 万幅各类场景图像和 4 万多个标注了 16 个关键点的

人体样本(如图 5 所示),其中 2.9 万个样本作为训练集,1.1 万个样本作为测试集,覆盖了 410 种人类日常活动。MPII 数据集提供了丰富标注信息,在测试集中还包括了人体部分遮挡和头部姿态方向的标注。

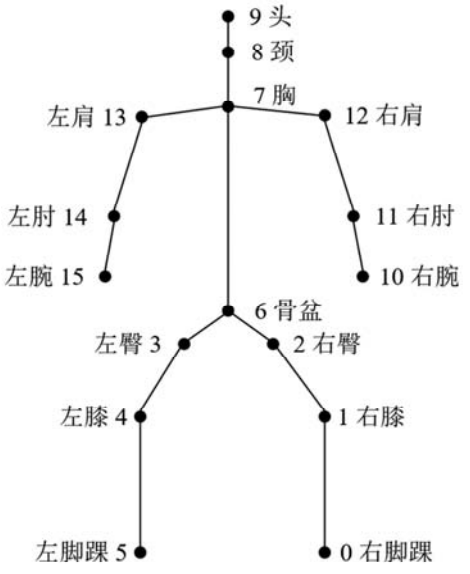


图 5　MPII 数据集骨骼关键点

2.2　实验环境

本文使用配置为 Intel Xeon E5-2623v4 CPU,NVIDIA Quadro P5000(16 GB 显存) GPU,128 GB 内存的高性能工作站对教师网络进行预训练。然后使用配置为 Intel Xeon E5-2650v4 CPU,NVIDIA Quadro M4000(8 GB 显存)GPU,64 GB 内存的工作站对学生网络进行训练和测试。

在训练过程中,首先对输入图像的特征进行压缩,特征大小压缩为 128。教师网络和学生网络在训练时的 batchsize 都等于 6,学习率为 2.6×10^{-4}。本文使用 RMSprop 优化器来避免损失函数更新过程中震荡幅度过大的问题。

本文算法采用关键点正确估计比例(Percentage of Correct Keypoints,PCK)作为评估指标。PCK 定义为如果关键点检测值与真实值之间的归一化距离小于设定阈值,则认为检测正确,计算公式如下:

$$PCK_{\delta}^{p}(T_0):=\frac{1}{|\alpha|}\sum_{\alpha}U(\|m_p^f-n_p^f\|_2<\delta) \quad (9)$$

式中:T_0 为被评估的关键点检测器,δ 为骨骼关键点的匹配阈值,α 为测试样本数量,m_p^f 表示预测值,n_p^f 表示真实值。

对于 MPII 数据集,本文的评估算法采用头部长度归一化,δ 取值为 0.5,即 PCKh@0.5。

2.3　实验结果与分析

在 MPII 数据集上,将本文方法测试结果与已有方法进行对比,具体指标包括模型参数数量、训练计算成本、人体各部位关键点检测准确率和平均准确率,测试结果如表 2 所示。其中,Teacher_n、LSHM_n、LSHM_t_n 分别表示教师网络模型,未经教师网络教导的模型和经过教师网络教导后的模型检测,n 表示级联的沙漏模型个数。

表 2　LSHM 模型检测结果对比

方法	参数数量/M	计算成本/GB	头	肩	肘	手腕	臀	膝盖	脚踝	均值
Wei[11]	31	351	97.8	95.0	88.7	84.0	88.4	82.8	79.4	88.5
Rafi[12]	56	28	97.2	93.9	86.4	81.3	86.8	80.6	73.4	86.3
Belagiannis[13]	17	95	97.7	95.0	88.2	83.0	87.9	82.6	78.4	88.1
Insafutdinov[14]	66	286	96.8	95.2	89.3	88.4	82.8	79.4	78.0	88.5
Ning[15]	74	124	98.1	96.3	92.2	87.8	90.6	87.6	82.7	91.2
Bulat[16]	76	67	97.9	95.1	89.9	85.3	89.4	85.7	81.7	89.7
Sekii[17]	16	6	—	—	—	—	—	—	—	88.1
Teacher_8[9]	26	12	96.6	95.5	89.3	84.5	88.7	84.2	80.1	88.7
LSHM_2	2.3	2.9	95.8	93.6	86.3	80.7	87.0	79.2	74.8	85.7
LSHM_4	12	3.6	96.2	95.2	88.6	83.2	88.1	83.1	79.1	88.0
LSHM_t_2	2.3	3	96.1	94.6	87.6	81.8	87.2	80.8	76.8	86.5
LSHM_t_4	12	5.4	96.4	95.4	89.0	84.1	88.6	83.9	79.6	88.5

相较于以往人体关键点检测网络,本文提出的LSHM 模型不仅提高了关键点的检测准确率,而且有效降低了参数规模和计算开销。对比未经过知识蒸馏的检测模型,经过知识蒸馏的模型可以获取教师网络学习后的知识,从而有效提高模型的检测能力。学生网络模型 LSHM_t4 仅比教师网络模型 Teacher_s8 的准确率降低了 0.2%,但 LSHM_t4 的参数总数仅为 Teacher_s8 的 46.1%,计算成本也仅为 Teacher_s8 的 45%,计算资源的消耗大大降低。

图 6(a) -(d)为本文算法的实际检测效果,可以看到本文所提出的改进轻量级沙漏模型可以正确检测出目标人物的骨骼关键点。特别是当出现图 6(c)中的肢体间遮挡与图 6(d)中的障碍物遮挡情况时,依然能有效获取人体骨骼关键点。

(a) (b) (c) (d)

图 6 本文算法关键点检测效果

3 结 语

本文通过引入逆残差卷积构建了改进轻量级沙漏模型,使用多尺度特征融合以提高轻量级模型在遮挡情况下的关键点检测能力。同时通过引入知识蒸馏方法,进一步提高了轻量级沙漏模型的关键点检测能力。MPII 数据集测试结果与实际检测效果表明本文提出的改进轻量级沙漏模型具有较好的关键点检测性能并能在一定程度上克服遮挡干扰,具有较好的实时性和较强的鲁棒性。

参 考 文 献

[1] Andriluka M, Pishchulin L, Gehler P, et al. 2D human pose estimation: New benchmark and state of the art analysis [C]//IEEE Conference on Computer Vision and Pattern Recognition,2014:3686 -3693.

[2] Chen Y C, Tian Y L, He M Y. Monocular human pose estimation: A survey of deep learning-based methods[J]. Computer Vision and Image Understanding,2020,192:102897.

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

[4] Sun K, Xiao B, Liu D, et al. Deep high-resolution representation learning for human pose estimation [C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition,2019:5693 -5703.

[5] Ke L P, Chang M C, Qi H G, et al. Multi-scale structure-aware network for human pose estimation [C]//European Conference on Computer Vision,2018:713 -728.

[6] Howard A G, Zhu M L, Chen B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications[EB]. arXiv:1704.04861,2017.

[7] Sandler M, Howard A, Zhu M L, et al. Mobilenetv2: Inverted residuals and linear bottlenecks [C]//IEEE Conference on Computer Vision and Pattern Recognition, 2018: 4510 -4520.

[8] Cao Z, Hidalgo G, Simon T, et al. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2019,43(1):172 -186.

[9] Luo D L, Du S L, Ikenaga T. End-to-end feature pyramid network for real-time multi-person pose estimation[C]//16th International Conference on Machine Vision Applications, 2019:1 -4.

[10] Zhang F, Zhu X T, Ye M. Fast human pose estimation [C]//IEEE Conference on Computer Vision and Pattern Recognition,2019:3517 -3526.

[11] Wei S E, Ramakrishna V, Kanade T, et al. Convolutional pose machines[C]//IEEE Conference on Computer Vision and Pattern Recognition,2016:4724 -4732.

[12] Rafi U, Leibe B, Gall J, et al. An efficient convolutional network for human pose estimation[C]//British Machine Vision Conference,2016:12.

[13] Belagiannis V, Zisserman A. Recurrent human pose estimation[C]//12th IEEE International Conference on Automatic Face & Gesture Recognition,2017:468 -475.

[14] Insafutdinov E, Pishchulin L, Andres B, et al. DeeperCut: A deeper, stronger, and faster multi-person pose estimation model[C]//European Conference on Computer Vision,2016: 34 -50.

[15] Ning G H, Zhang Z, He Z Q. Knowledge-guided deep fractal neural networks for human pose estimation[J]. IEEE Transactions on Multimedia,2017,20(5):1246 -1259.

[16] Bulat A, Tzimiropoulos G. Human pose estimation via convolutional part heatmap regression[C]//European Conference on Computer Vision,2016:717 -732.

[17] Sekii T. Pose proposal networks[C]//European Conference on Computer Vision,2018:350 -366.