第 58 卷 第 20 期/2021 年 10 月/激光与光电子学进展 研究论文

激光与光电子学进展 先进成像

基于改进沙漏网络的人体姿态估计模型

刘红，马杰*，柴玉晶

河北工业大学电子信息工程学院，天津 300401

摘要 堆栈沙漏网络(SHN)是人体姿态估计中的代表性研究成果，但该网络忽略了关节局部信息。因此，提出了一种基于改进沙漏网络的人体姿态估计模型。首先，利用多个残差模块及步长为 2 的卷积层获取低层次到高层次的特征，同时随着网络层数的加深，相应调整残差模块的数目和通道数，以突出局部细节特征信息。然后，为了提取遮挡部位的纹理和形状等局部特征，融合了在线困难关键点挖掘模块。最后，采用反卷积最大化恢复原始的局部特征。实验结果表明，本模型在 COCO 数据集上的平均精度达到了 74.6%，总参数量为 1.5×10^7，比叠加 8 个 SHN(8-SNH)的平均精度高 5.1 个百分点，且其总参数量仅为 8-SNH 的 1/3。

关键词 机器视觉；人体姿态估计；堆栈沙漏网络；残差模块；在线困难关键点挖掘

中图分类号 TP391 **文献标志码** A **doi**：10.3788/LOP202158.2015004

Human Pose Estimation Model Based on Improved Hourglass Network

Liu Hong, Ma Jie*, Chai Yujing

School of Electronics and Information Engineering, Hebei University of Technology, Tianjin 300401, *China*

Abstract A stacked hourglass network (SHN) is a representative research result in human pose estimation; however, it ignores the local information of joints. Therefore, this study proposes a human pose estimation model based on an improved hourglass network. First, multiple residual modules and convolution layer with a step size of 2 are used to obtain low- to high-level features. In order to highlight the local detailed feature information, the number of residual modules and channels are adjusted as the number of network layers deepens. Then, an online difficult keypoint mining module is integrated to extract local features such as texture and shape of the occluded part. Finally, deconvolution is used to maximize the restoration of the original local features. The experimental results show that the average accuracy of the proposed model on the COCO data set reaches 74.6%. In addition, the total parameter amount is 1.5×10^7, which is 5.1 percentage points higher than the average accuracy of superimposing eight SHN (8-SHN), and its total parameter amount is only 1/3 of 8-SHN.

Key words machine vision; human pose estimation; stacked hourglass network; residual module; online hard keypoint mining

OCIS codes 150.4065; 100.4996; 150.1135

1 引 言

人体姿态估计是计算机视觉领域的一项重要任务，其目标是从复杂场景中检测人体关节的位置，并输出全部或局部关节的信息[1-3]，被人们广泛应用于虚拟现实(VR)电影、人机交互、智能监控等领域中[4-6]。已有的姿态估计方法主要存在两个难点，一方面，姿态估计任务本身存在拍摄角度差异、背景混淆、姿态差异等难点，导致图片中的人体存在不同程度的遮挡；另一方面，姿态估计方法主要关注检测精

收稿日期：2020-12-02；**修回日期**：2020-12-28；**录用日期**：2021-01-13

基金项目：河北省自然科学基金(F2020202045)、河北省研究生创新项目(CXZZBS2020026)、天津市教委科研计划(2018KJ268)

通信作者：*jma@hebut.edu.cn

度，而忽略了模型的运算速率。

传统姿态估计方法往往依靠梯度方向直方图（HOG）[7]和尺度不变特征变换（SHIFT）[8]等特征，虽在特定场景下能达到较高的检测速度与精度，但这类方法依赖先验知识，模型的自适应性及泛化性较差[9-10]。而基于深度学习的方法摆脱了对上述条件的依赖，可自适应提取各部件的特征，并将训练好的模型应用在不同的场景中，有效提高了检测精度。Toshev 等[11]将深度神经网络（DNN）应用于人体关节的识别，用 AlexNet 捕捉不同关节的高低分辨率特征，并通过全连接层直接回归关节的坐标点，但该方法获得的权重依赖于训练数据的分布。Wei 等[12]提出的卷积姿态机（CPM）是一种多阶段级联顺序卷积结构，在每个阶段均加入监督训练，以各部件的热力图表达各部件之间的空间约束，但该结构的实现难度较大。因此，Newell 等[13]在文献[12]的基础上提出了堆栈沙漏网络（SHN），该网络由若干个沙漏网络（HN）叠加而成，每个阶段均采用热力图计算损失，并借助最大池化和近邻域插值融合网络低层次和高层次的特征图（Feature map）。但该网络中用堆栈定义多个 HN，需要消耗大量的计算资源；且网络简单的叠加多个上采样、下采样和残差模块，丢失了大量的局部特征信息，不利于提取关节部位的纹理和形状等特征。此外，随着训练时间的增加，SHN 会倾向于简单点（容易识别的关节），而忽略困难点（遮挡的关节），不利于困难点位置信息的提取。

针对上述问题，本文提出了一种改进的沙漏网络（IHN）。首先，用一个 IHN 取代叠加的 8 个 HN，在保证模型性能的同时，提高了网络的运行速率。然后，构建了一个预测模块，该模块每层采用若干个残差模块且只含有一次步长为 2 的卷积操作，可根据网络的层数调整残差模块的层数和通道数，使网络在保留从局部到整体结构化信息的同时，加强网络对局部细节特征的学习能力。最后，用改进的反卷积模块扩大关节的特征图，直接用反卷积层取代下采样和残差模块，从而更细致地恢复原始图像的像素。针对网络训练过程中简单点和困难点不平衡的问题，构建出在线困难关键点挖掘（OHKM）模块，加强了网络对困难点部位纹理和形状等特征的学习。

2　算法描述

2.1　沙漏网络

HN 呈对称状结构且形似沙漏，如图 1 所示，其中，每个方框表示该阶段各关节的特征图。在 HN 的前半部分，每层网络通过 3 个残差模块[14]及下采样[15]操作得到分辨率逐渐降低的特征图，并向后传递到网络中心，得到分辨率最低的特征图；后半部分，每层网络通过上采样[16]和 1 个残差模块逐步恢复出高分辨率的特征图。同时，HN 跳跃层经过逐步提取向后半部分网络传递关节特征。最后，将跳跃层保留的各尺度特征与后半部分得到的低分辨率特征进行融合。HN 通过自下而上和自上而下的结构将网络低层次和高层次的特征图进行融合，以捕捉人体各关节的空间位置信息。

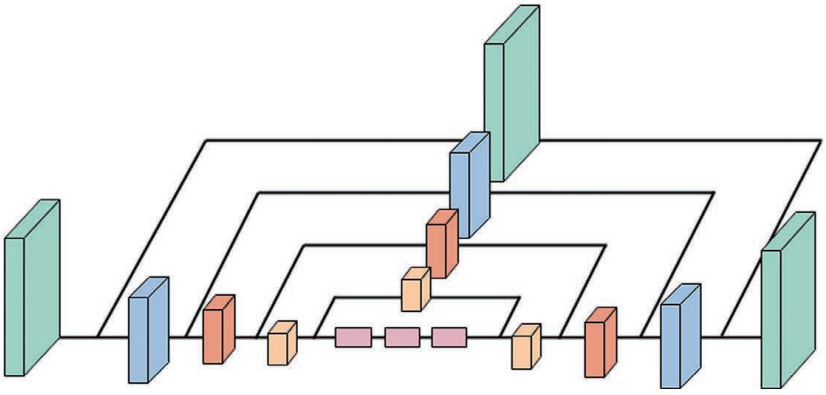

图 1　HN 模型的结构

Fig. 1　Structure of the HN model

2.2　基于改进沙漏网络的人体姿态估计方法

对于需要检测 K 个关节点的任务，$\boldsymbol{P}_K \in \boldsymbol{R}^{W\times H\times M}$ 为第 K 个关节点的位置，$W\times H\times M$ 为图像尺寸，IHN 的任务是从图像中找出各个关节点 $\{\boldsymbol{P}_1,\cdots,\boldsymbol{P}_K\}$ 的位置。图 2 为 IHN 的具体结构，主要包括 4 个模块。其中，N1 为预测模块，可通过各关节在不同尺度下的特征图预测出各关节的位置；N2 为 OHKM 模块，可判别各关节检测的难易程

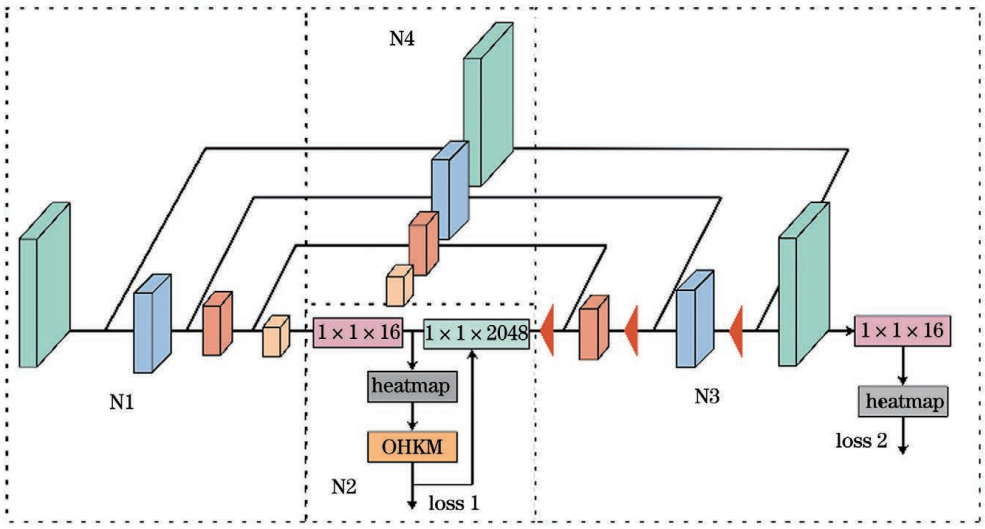

图 2　IHN 模型的结构

Fig. 2　Structure of the IHN module

度，增强网络对困难点的泛化能力；N3 为反卷积层模块，通过不断优化预测结果精修各关节的位置；N4 为跳跃层模块，可保留各尺度下关节的原始特征，并通过反卷积模块继续进行信息融合。

整个检测流程中，首先，用由高分辨率到低分辨率的特征图预测各关节的位置。然后，将得到的各关节信息以热力图的形式传入 OHKM 模块中，以判别各关节检测的难易程度，并对检测出的困难点进行反向传播，用于更新模型的参数。在后续工作中，将高倍降采样的特征图传入反卷积模块，并通过反卷积叠加到相应尺度的特征图中。最后，通过迭代优化检测结果，从低分辨率到高分辨率的特征图上得到最优的各关节信息。

2.2.1　预测模块

简单的池化会丢失大量特征信息，为了更细致地学习关节特征，提出了一种预测模块，其结构如图 3 所示。其中，C_i 为第 i 层预测模块的输出特征。每个方块均表示残差模块。每层预测模块采用若干

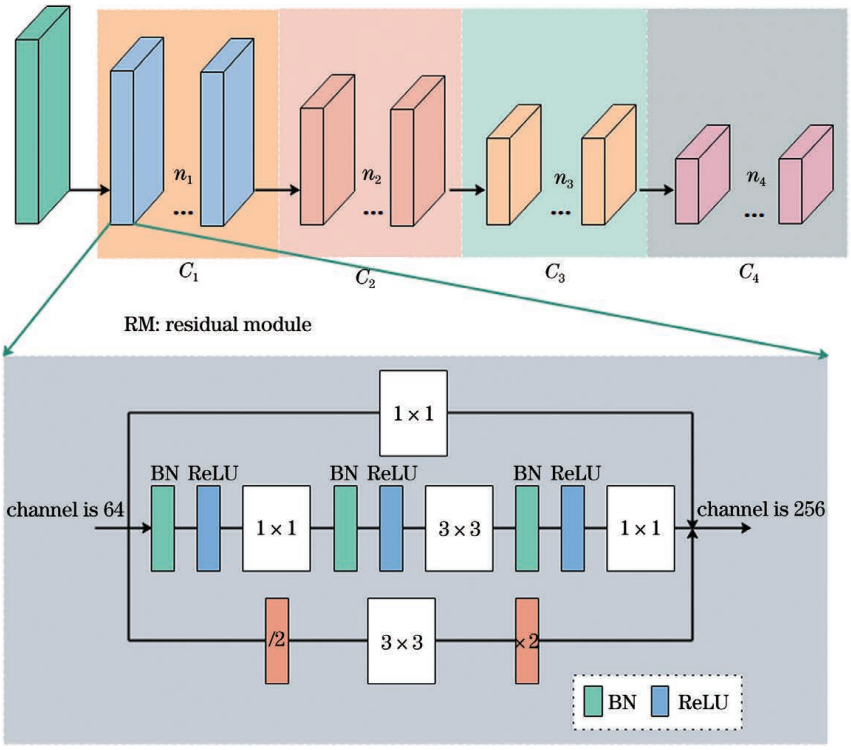

图 3　预测模块的结构

Fig. 3　Structure of the prediction model

个残差模块且只含有一次步长为 2 的卷积操作。从原始图像中进行关节检测需要消耗大量的内存，因此，先通过 64 个尺寸为 7×7、步长为 2 的卷积核对图像进行预处理。随着网络从低层到高层的进行，提取的特征逐渐呈从整体到局部的信息。为了使关节特征更好地适应网络层数的变化，突出局部细节特征信息，还需相应调整残差模块的数目和通道数。

图 3 中，n_i 为第 i 层预测模块中总残差模块的数量，可表示为

$$n_i = i \times 2 + 1, \tag{1}$$

预测模块中的总残差模块数量为

$$n = \sum_{i=1}^{i} n_i 。 \tag{2}$$

若输入图像为 $\boldsymbol{X} \in \boldsymbol{R}^{W \times H \times M}$，则第 n 层残差模块输出的特征可表示为

$$f_n'(\boldsymbol{X}^n; w_{f_n}^n) = \sum_{n=1}^{n} \left[f_n(\boldsymbol{X}^n; w_{f_n}^n) + w_s^n h(\boldsymbol{X}^n) \right], \tag{3}$$

式中，$\boldsymbol{X}^n$ 为第 n 层的输入特征，$h(\cdot)$ 为特征映射，$f_n(\cdot)$ 为第 n 层残差模块特征的变换函数，$w_{f_n}^n$ 为第 n 层残差模块的权值，w_s 为匹维度。第 i 层预测模块的输出特征可表示为

$$P_{n_i}(\boldsymbol{X}^{n_i}; w^{n_i}) = g_{n_i}\left[\sum_{n_1}^{n_i} f_n'(\boldsymbol{X}^n; w_{f_n}^n); w_g^n \right], \tag{4}$$

式中，$g(\cdot)$ 为卷积，w_g^n 为第 n 层卷积的权值。

2.2.2　在线困难关键点挖掘模块

IHN 虽然可以提高模型的检测精度，但仍然存在缺陷，如在关节重叠、关节不可见以及背景拥挤导致部分关节被遮挡的情况下很难实现定位。因此，对在线样本挖掘（OSHM）模块进行改进，构造出了一种适用于人体姿态估计的 OHKM 模块，解决了困难点的检测问题。OHKM 模块的流程如下。

1）经过 1×1×17 的瓶颈网络，将网络通道变为热力图的接收模式。

2）利用二维（2D）高斯函数生成 K 个关节点的热力图，若关节点的坐标为 (x, y)，则对应的热力图转换公式为

$$F(x, y) = X_{\text{size}} \times \exp\left[-\frac{(x - x_0)^2 + (y - y_0)^2}{\alpha^2} \right], \tag{5}$$

$$X_{\text{HM}}(w, h, k)_{0<k<K} = F(x_k, y_k)_{0<x_k<W, 0<y_k<H}, \tag{6}$$

式中，X_{size} 为高斯函数的尺寸（幅值），α^2 为高斯函数的方差，(x_0, y_0) 为人体关节点坐标，$X_{\text{HM}}(w, h, k)_{0<k<K}$ 为第 k 个关节点的热力图，(x_k, y_k) 为根据标准转换成的关节点坐标。

3）由热力图的输出值计算损失 L_1[17]，可表示为

$$L_1 = \sum_{p=1}^{P} \sum_{k=1}^{K} \left\| b^p(k) - b_*^p(k) \right\|^2, \tag{7}$$

式中，P 为图像中人的数量，b_* 为 Ground truth，b 为网络的预测值。

4）L_1 可表示模型对当前每个关节点的检测性能，因此，将 L_1 按从大到小的顺序排列，取性能最差（前 M 个）的关节点进行反向传播，用于更新模型的参数，从而增强网络对困难点的泛化能力。

5）经过 1×1×1024 的瓶颈网络，整合各关节点的通道数，以便后续的反卷积模块更细致地学习各关节点特征。

2.2.3　反卷积模块

HN 通过上采样扩大分辨率，将输入图像恢复到一个理想尺寸，并计算每个点的像素值，再使用近邻域插值方法对周围像素点进行插值处理，具体步骤如图 4(a)、图 4(b)所示。但该方法得到的高分辨率特征图与周围像素值的区分度不够（周围像素点的值全为 0 或全为相同值），导致恢复的特征图与原始图像（原始图像的周围像素值大部分不相同）的差异很大。反卷积操作能使周围像素得到不同的值，恢复出与原始图像差异较小的图像，如图 4(d)所示。虽然该方法不能还原出原始信息，但可以最大化恢复缺失的局部信息，从而使模型在匹配人体关节时的损失更小、精度更高。

HN 先通过上采样扩大分辨率，再用残差模块提取关节特征。而本方法直接用反卷积替代上采样和残差模块，反卷积模块可被细分成 3 层反卷积层，每层反卷积的操作大体一致。若反卷积层输入的图像分辨率为 w，则输出图像的分辨率可表示为

$$w_o = X_s \times (w - 1) + X_k - 2X_p, \tag{8}$$

式中，X_s 为步长，X_k 为卷积核的尺寸，X_p 为填充因子。

反卷积模块可以根据预测模块特征图的尺度适当调整卷积核和通道数，预测模块采用4组残

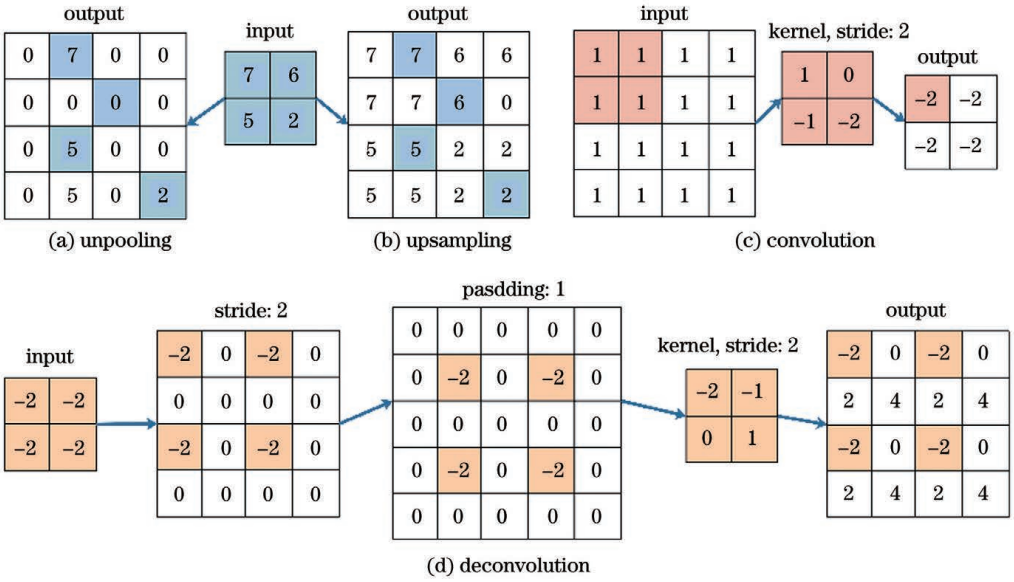

图 4　扩大分辨率的原理。(a)反池化;(b)上采样;(c)卷积;(d)反卷积

Fig. 4　Principle of expanding resolution. (a) Unpooling; (b) upsampling; (c) convolution; (d) deconvolution

差模块，由(3)式得到预测模块的输出特征为 $P_{n_4}[x^{(n_4)};w^{(n_4)}]$，则该模块第 d 层的输出特征为

$$P_{d_i}[x^{(d_i)};w^{(d_i)}]=g_d\left\{\sum_{j=1}^{D}P_{n_4}[x^{(n_4)};w^{(n_4)}];w_d^{(j)}\right\}+P[x^{(n_i)};w^{(n_i)}],\tag{9}$$

式中，$d+i=5$，$g_d(\cdot)$为反卷积操作，$P_H[x^{(n_4)};w^{(n_4)}]$为经过 OHKM 模块处理后的特征，$w_d^{(j)}$ 为第 d 层的反卷积权值，i 为反卷积的层数。

3　实验设置与结果分析

为评估 IHN 的性能，在 COCO 数据集[18]和 MPII 数据集上进行训练和测试，并将本方法与相关方法得到的实验结果进行对比。实验环境：显卡为 GTX 1080Ti，操作系统为 Ubuntu16.04 LTS，软件包括 Python(3.6 版)、TensorFlow(1.12.0 版)，人体检测器为区域卷积神经网络(RCNN)，优化器为 Adam 算法。

3.1　实验数据集及评估标准

通过两个公开基准数据集(COCO 数据集和 MPII 数据集)对本方法进行训练和测试。COCO 数据集对应网络的输入尺寸为 256×192，MPII 数据集对应网络的输入尺寸为 256×256。为了对数据进行标准化，减小误差，设图像预处理参数在各维度的均值 $\boldsymbol{M}=[0.485,0.456,0.406]$、标准差 $\boldsymbol{S}=[0.229,0.224,0.225]$，以人体为中心采用随机旋转[−40°,40°]、缩放[−30%,30%]和翻转等方式对数据集进行扩增处理。COCO 数据集的评价指标为基于关键点相似性(OKS)的平均精确度(AP)和平均召回率(AR)；MPII 数据集的评价指标为关键点准确估计百分比(PCK)。

3.2　实验结果及分析

用 DeepPose[11]、CPM[12]、8-SHN[13]、Baseline[19]方法及 IHN-OHKM 方法(本方法)在 COCO 测试集上对各关节点进行测试，输出结果如图 5 所示。可以发现，采用坐标回归的 DeepPose[11]及利用多阶段级联的 CPM 方法未能估计出完整体态，8-SHN[13]和 Baseline[19]方法未能准确检测出图像中人体的腕部、肘部、眼部和鼻子等，而本方法在复杂环境下对困难点部位的检测效果较好，整体效果优于其他方法，如本方法可以很好地解决图 5(a)中左侧人右手腕被遮挡的问题。

本方法采用的 Batchsize 为 128、Epochs 为 120、Iteration 为 2000。由于学习率过大会导致损失振荡，因此，网络设置的初始学习率为 10^{-4}，当 Epochs 达到 80 后，将学习率降为 10^{-5}。不同方法在 COCO 数据集上的 AP 曲线如图 6 所示，可以发现，当 Epochs 达到 80 后，不同方法的性能逐渐趋于稳定，但本方法能达到的 AP 更高且趋于稳定的时间更短。

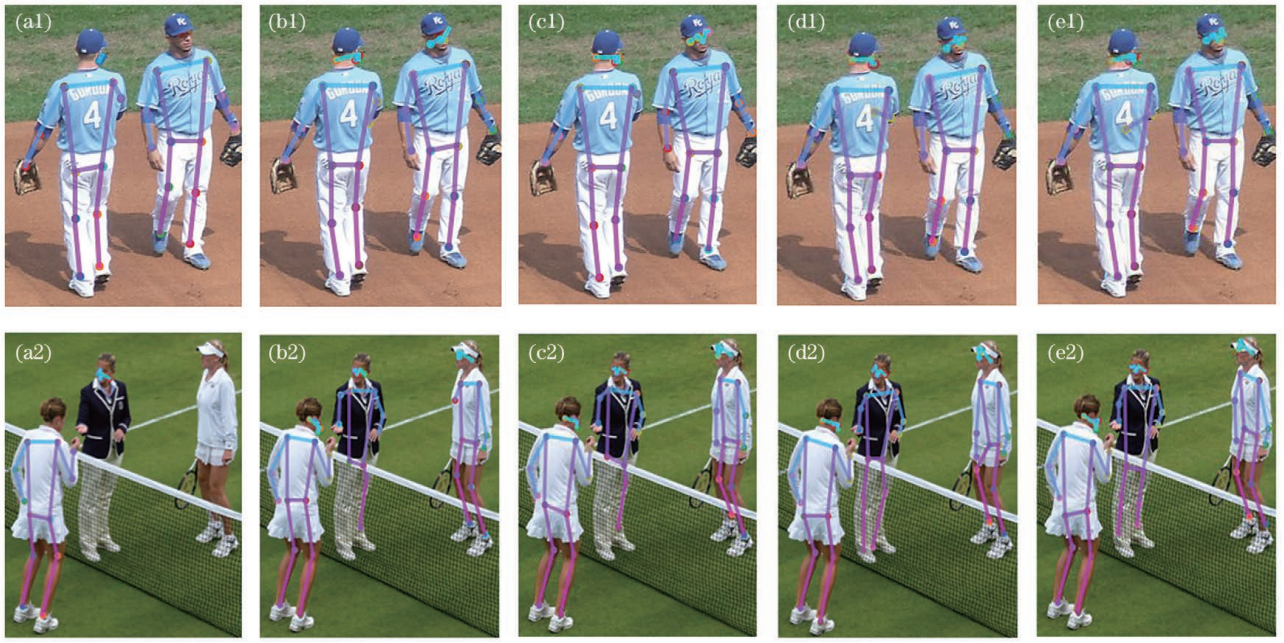

图 5　不同方法在 COCO 数据集上的实验结果。(a)DeepPose[11]；(b)CPM[12]；(c)8-SHN[13]；(d)Baseline[19]；(d)IHN-OHKM

Fig. 5　Experimental results of different methods on the COCO data set. (a) DeepPose[11]; (b) CPM[12]; (c) 8-SHN[13]; (d) Baseline[19]; (e) IHN-OHKM

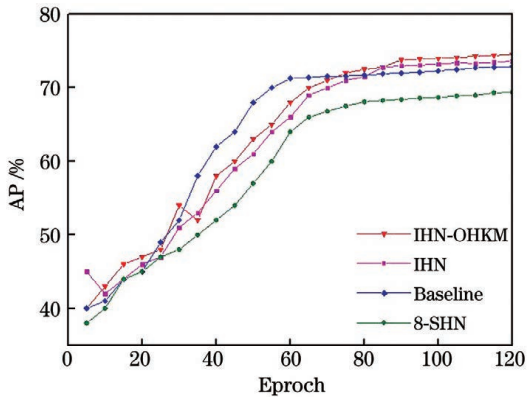

图 6　不同方法在 COCO 数据集上的 AP 曲线

Fig. 6　AP curves for different methods in COCO data set

为了评估本方法对多人姿态估计的性能，对比了本方法与其他相关方法的准确率，结果如表1所示。可以发现，相比 8-SHN 方法，本方法的 AP 和 AR 分别提升了 5.1 和 2.8 个百分点。相比 DeepPose 方法，本方法的 AP 和 AR 分别提升了 8.0 和 7.7 个百分点，原因是 DeepPose 方法依赖于坐标回归，学习的滤波器以粗略的比例捕获姿态属性，不能精确定位身体关节，而本方法直接采用网络生成的热力图，减少了网络坐标转换带来的误差。相比 Baseline 方法，本方法的 AP 和 AR 分别提升了 1.6 和 1.4 个百分点。原因是 Baseline 方法忽略了高层次特征与低层次特征的融合问题，无法充分利用所有尺寸的特征图。加入 OHKM 模块后，本方法的 AP 和 AR 分别达到了 74.5% 和 79.6%。其中，$AP^{@50}$ 表示关节相似度大于等于 50%；$AP^{@m}$ 为中等目标的平均精度，$AP^{@l}$ 为大目标的平均精度。

表 1　不同方法在 COCO 数据集上的检测结果

Table 1　Test results of different methods on the COCO data set　　unit：%

Method	AP	$AP^{@50}$	$AP^{@75}$	$AP^{@m}$	$AP^{@l}$	AR
DeepPose[11]	66.5	80.6	73.6	63.4	72.7	71.9
CPM[12]	66.9	82.3	76.3	65.3	75.9	74.4
8-SHN[13]	69.4	85.3	79.8	67.4	78.6	76.8
Baseline[19]	72.9	88.5	80.2	69.0	**79.3**	78.2
IHN	73.6	88.7	**80.5**	69.1	79.0	79.1
IHN-OHKM	**74.5**	**88.8**	80.2	**69.4**	78.5	**79.6**

在相同情况下，对比了本方法与 IHN、8-SHN 方法的效率，包括平均处理时间、模型浮点运算次数（GFLOPs）及模型参数数量，结果如表 2 所示。可以发现，相同迭代次数下，本方法的参数量约为 8-SHN 方法的 1/3，大幅度简化了模型的复杂度；且本方法处理一张图像的平均用时约为 23 ms，满足实时性检测的要求。相比 IHN，融合 OHKM 的 IHN 虽然检测精度较高，但同时伴随着使用效率降低和复杂度增加的问题。

表 2　不同方法的计算效率

Table 2　Calculation efficiency of different methods

Method	Average processing time /ms	GFLOPs /(10^9 times)	Number of parameters
8-SHN[13]	66	20.3	4.6×10^7
IHN	**20**	**6.4**	**1.3×10^7**
IHN-OHKM	23	6.9	1.5×10^7

图 7 为本方法在 COCO 数据集上的检测结果，图中每一个人均存在不同程度的遮挡。可以发现，本方法能准确估计各关节的位置，对多人姿态估计的关节定位效果较好。

图 7　本方法在 COCO 数据集上的检测结果

Fig. 7　Detection results of our method on the COCO data set

MPII 测试集中的数据包含 7 种主要关节，分别为头部（Head）、肩部（Sho.）、肘部（Elb.）、腕部（Wri.）、髋部（Hip）、膝部（Knee）、踝部（Ank.）。为了评估不同方法的单人姿态估计性能，将本方法与其他方法的实验结果进行对比，结果如表 3 所示。可以发现，8-SHN[13] 方法的检测效果较好，IHN-OHKM 的平均准确率达到了 92.9%。相比 IHN，融合 OHKM 模块后，网络对于“困难点”（髋部）的平均准确率提升了 1.4 个百分点。其中，阈值 r（关节相似度）为 0.5。

表 3　不同方法在 MPII 数据集上的检测结果

Table 3　Test results of different methods on MPII data set　　unit：%

Method	Head	Sho.	Elb.	Wri.	Hip.	Knee	Ank.	Mean
DeepPose[11]	95.4	94.3	91.7	84.0	89.7	87.0	81.3	89.1
CPM[12]	96.2	95.0	92.0	84.9	90.5	87.7	82.0	89.8
8-SHN[13]	**97.6**	95.7	92.3	85.3	91.3	88.6	82.5	90.5
Baseline[19]	97.0	96.5	93.4	88.5	92.0	**90.7**	83.2	91.8
IHN	97.3	97.8	94.2	89.0	92.6	90.5	85.9	92.5
IHN-OHKM	97.0	**98.9**	**95.3**	88.6	**94.0**	90.2	**86.4**	**92.9**

3.3 消融实验

在 COCO 验证数据集下，对不同预测模块和反卷积的层数进行了消融实验，结果如表 4 所示。为了便于跳跃层和反卷积模块的特征融合，使后续人体关节检测匹配率最高，反卷积模块的层数需要与预测模块相对应。实验对比了使用 3、4、5 层预测模块的性能，对应的反卷积层分别为 2、3、4 层。可以发现，随着网络层数的加深，网络的性能有所提高，但复杂度也逐渐增大，因此，实验采用 4 层预测模块和 3 层反卷积层。

表 4 COCO 验证数据集上的消融实验

Table 4 Ablation experiments on the COCO validation data set

Residual module layer	Deconvolution layer	AP /%	Average processing time /ms
3	2	71.2	16.6
4	**3**	**74.6**	**18.0**
5	4	75.0	21.4

在 COCO 验证数据集下，选取损失最大的 M ($M < K=17$)个关节进行反向传播，M 的取值对实验结果的影响如表 5 所示。可以发现，M 的取值并非越大越好。对于较大的 M，网络的分类性能会下降。当 $M=8$ 时，网络的 AP 达到最高，为 75.5%。

表 5 不同 M 对 AP 的影响

Table 5 Impact of different M on AP

M	6	**8**	10	12	14	17
AP /%	70.4	**75.5**	74.1	73.0	72.7	72.5

4 结 论

在传统 HN 的基础上进行改进，设计并实现了一种基于 IHN 的人体姿态估计网络。该网络包含预测模块、OHKM 模块、跳跃层模块及反卷积模块。预测模块利用残差模块以及步长为 2 的卷积层提取网络不同层次的特征图，同时针对预测模块的不同层调整残差模块的数目和通道数，增加网络对更多局部细节信息的聚焦。OHKM 模块通过识别出的困难点更新模型参数，增强网络对困难点纹理等局部特征的学习。反卷积模块在扩大分辨率的同时使中心像素周围的像素值多样化，从而得到更接近原始图像的特征图。实验结果表明，相比 8-SHN 方法，本方法的模型精度高、参数量少，且能有效识别出被遮挡的关节，但如何更好地利用关键点检测技术进行人体行为分析、姿态预测，仍是亟需解决的问题。

参 考 文 献

[1] Aggarwal J K, Ryoo M S. Human activity analysis: a review[J]. ACM Computing Surveys, 2011, 43(3): 1-43.

[2] Datta R, Joshi D, Li J, et al. Image retrieval, ideas, influences, and trends of the new age[J]. ACM Computing Surveys, 2008, 40(2): 1-60.

[3] Palmese M, Trucco A. From 3-D sonar images to augmented reality models for objects buried on the seafloor[J]. IEEE Transactions on Instrumentation and Measurement, 2008, 57(4): 820-828.

[4] Chen Y, An W Y, Liu H L, et al. Application of improved empirical mode decomposition algorithm in fiber Bragg grating perimeter intrusion behaviors classification[J]. Chinese Journal of Lasers, 2019, 46(3): 0304003.
陈勇, 安汪悦, 刘焕淋, 等. 改进经验模态分解算法在光纤布拉格光栅周界入侵行为分类中的应用[J]. 中国激光, 2019, 46(3): 0304003.

[5] Bi X J, Wang H. Person re-identification based on view information embedding[J]. Acta Optica Sinica, 2019, 39(6): 0615007.
毕晓君, 汪灏. 基于视角信息嵌入的行人重识别[J]. 光学学报, 2019, 39(6): 0615007.

[6] Liu F, Yu F Q. Human action recognition based on global and local features[J]. Laser & Optoelectronics Progress, 2020, 57(2): 021004.
刘帆, 于凤芹. 基于全局和局部特征的人体行为识别[J]. 激光与光电子学进展, 2020, 57(2): 021004.

[7] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]// 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), June 20-25, 2005, San Diego, CA, USA. New York: IEEE Press, 2005: 886-893.

[8] Lowe D G. Distinctive image features from scale-invariant keypoints[J]. International Journal of Computer Vision, 2004, 60(2): 91-110.

[9] Andriluka M, Roth S, Schiele B. Pictorial structures revisited: people detection and articulated pose estimation[C]//2009 IEEE Conference on Computer Visionand Pattern Recognition, June 20-25, 2009, Miami, FL, USA. New York: IEEE Press, 2009: 1014-1021.

[10] Ladický L, Torr P H S, Zisserman A. Human pose estimation using a joint pixel-wise and part-wise formulation[C]//2013 IEEE Conference on Computer

Vision and Pattern Recognition, June 23-28, 2013, Portland, OR, USA. New York: IEEE Press, 2013: 3578-3585.

[11] Toshev A, Szegedy C. DeepPose: human pose estimation via deep neural networks[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2014, Columbus, OH, USA. New York: IEEE Press, 2014: 1653-1660.

[12] Wei S H, Ramakrishna V, Kanade T, et al. Convolutional pose machines[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 4724-4732.

[13] Newell A, Yang K Y, Deng J. Stacked hourglass networks for human pose estimation[M]//Leibe B, Matas J, Sebe N, et al. Computer vision-ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016, 9912: 483-499.

[14] He K M, Zhang X Y, Ren S Q, et al. Identity mappings in deep residual networks[M]//Leibe B, Matas J, Sebe N, et al. Computer vision-ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016, 9908: 630-645.

[15] Hang S T, Aono M. Bi-linearly weighted fractional max pooling[J]. Multimedia Tools and Applications, 2017, 76(21): 22095-22117.

[16] Olivier R, Cao H Q. Nearest neighbor value interpolation[J]. International Journal of Advanced Computer Science and Applications, 2012, 3(4): 25-30.

[17] Zhao H, Gallo O, Frosio I, et al. Loss functions for image restoration with neural networks[J]. IEEE Transactions on Computational Imaging, 2017, 3(1): 47-57.

[18] Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: common objects in context[M]//Fleet D, Pajdla T, Schiele B, et al. Computer vision-ECCV 2014. Lecture notes in computer science. Cham: Springer, 2014, 8693: 740-755.

[19] Xiao B, Wu H P, Wei Y C. Simple baselines for human pose estimation and tracking[M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11210: 472-487.