复杂运动场景下的多人姿态估计研究

柳长源*，臧彦丞，兰朝凤

（哈尔滨理工大学 测控技术与通信工程学院，哈尔滨 150080）

摘要 针对运动场景中运动员之间的相互遮挡、自身部位遮挡、运动器械遮挡及复杂背景干扰问题，提出一种高分辨特征生成复原网络，引入融合注意力机制筛选有用特征信息通道，加入反卷积和多尺度特征融合模块分层处理小目标人像与大中型目标人像的姿态估计任务。设计生成对抗模块，对缺失部分进行补全和预测得到关节点热图，经过姿态骨架和最优匹配算法确定出关节点连接方式，并输出可视化姿态估计结果。在 MSCOCO 和 Crowd Pose 数据集上的实验结果表明该姿态估计方法在复杂运动场景下效果更优。

关键词 人体姿态估计；深度学习；复杂运动场景；融合注意力机制；生成对抗网络

中图分类号 TP391.4 **文献标志码** A DOI 10.12178/1001-0548.2023233

Research on Multiplayer Pose Estimation in Complex Sports Scenes

LIU Changyuan*, ZANG Yancheng, and LAN Chaofeng

(College of Measurement and Control Technology and Communication Engineering, Harbin University of Science and Technology, Harbin 150080, China)

Abstract In terms of the problems such as mutual occlusion, self-occlusion, sports equipment occlusion and complex background interference among athletes in motion scenes, a high-resolution feature generation recovery network is proposed in this paper. The attention fusion mechanism is introduced to screen the useful feature information channels. The deconvolution and multi-scale feature fusion modules are added to deal with the pose estimation tasks for small target portraits and large and medium-sized target portraits in a hierarchical manner. The adversarial module is designed and generated to complete and predict the missing parts to obtain the keypoint heat map, the keypoint connection mode is determined through the pose skeleton and the optimal matching algorithm, and the visual pose estimation results are output. Experimental results on MSCOCO and Crowd Pose datasets have showed that the pose estimation method is more effective in complex motion scenes.

Key words human pose estimation; deep learning; complex motion scenes; fusion attention mechanisms; generate adversarial networks

人体姿态估计的研究对象是图像或视频中的人像，具体表现方式是对检测到的人像定位出关节点的正确位置并标注出来，最后可视化输出到结果图上。人体姿态估计被广泛应用于动作识别[1]、体育健身[2]等重要领域，在深度学习的发展和推动下不断丰富和完善。

传统的姿态估计方法依赖手工设计的特征[3]和人体模型[4]进行姿态估计。文献 [5] 提出了一种基于深度神经网络（Deep Neural Network, DNN）的人体姿态估计方法，将姿态估计问题转化为基于 DNN 的回归问题，整体推导人体姿态，实现高精度姿态估计。文献 [6] 受文献 [7] 启发提出了一种基于序列化的网络结构卷积位姿机（Convolutional Pose Machines, CPM），利用 CPM 学习丰富的隐式空间模型，能够有效预测被遮挡人体的关节点。文献 [8] 将多内容信息注意力机制整合到卷积神经网络（Convolution Neural Network, CNN）中，得到端到端的人体姿态估计框架，并选用堆叠沙漏网络（Stacked Hourglass Networks, SHN）形成不同分辨率的注意力图，再利用条件随机场（Conditional Random Field, CRF）对注意力图中相邻区域的关联性进行建模，从而提高了关节点定位的准确性[9]。

收稿日期：2023−09−21；修回日期：2024−01−15

基金项目：国家自然科学基金（11804068）；黑龙江省交通运输厅科技项目（HJK2024B002）

作者简介：柳长源，博士，副教授，主要从事图像处理、模式识别与机器学习方面的研究。

*通信作者 E-mail: liuchangyuan@hrbust.edu.cn

文献 [10] 提出了一种以残差网络为主网络的 PifPaf 算法和部位强度场（Part Intensity Field, PIF）的概念来解决低分辨率和拥挤场景下的姿态估计问题。文献 [11] 提出高分辨率网络（High Resolution Networks, HRNet），并行连接不同分辨率的卷积，以提升高分辨率的表示。文献 [12] 提出了 HigherHRNet 网络，该网络在 HRNet 的基础上引入高效的反卷积模块，并采用了多分辨率训练和热图聚合策略，提升了关节点的预测效率。文献 [13] 提出了深度卷积生成对抗模型（Deep Convolution Generative Adversarial Network, DCGAN）设计分层对抗网络（Hierarchical Adversarial Network, HAN）和层次感知损失，以提高身体各个部位的位置估计精度。文献 [14] 提出 Lite-HRNet 用于轻量化 HRNet 的网络结构，将 ShuffleNet 中的高效 Shuffle 块用于 HRNet，提升网络性能并达到轻量和高效的姿态估计效果。文献 [15] 提出 DEKR（Disentangled Keypoint Regression）算法，采用多分支结构，分支与关节点对应，然后利用自适应卷积激活关节点周围的像素，并学习关节点周围像素的特征，基于这些特征回归关节点的位置，实现各个关节点之间的解耦。文献 [16] 提出动态轻量化网络 Dite-HRNet，用动态拆分卷积和自适应上下文建模的方法解决高分辨率网络无法捕获大范围相互作用关节点的问题。文献 [17] 提出 HDFormer（High-Order Directed Transformer）框架，将自注意力、高阶注意力机制和卷积神经网络整合，从而降低模型的参数量。文献 [18] 提出 ED-Pose 方法，引入检测解码器来提取全局特征，将姿势估计视为一个关键点盒子检测问题，学习每个关键点的盒子位置和内容，无须后期处理和密集的热图监督。

综合上述的研究方法发现：传统的姿态估计方法需要人工干预，算法效率受限于复杂环境因素，在遮挡场景和复杂背景干扰情况下精度偏低，在遮挡和光照不均的情况下，容易漏检或误检人体关节点，相同人体的不同姿态的预测精度也存在差异。本文基于上述复杂运动场景的难点问题提出高分辨特征生成复原网络，该网络由高分辨特征图网络和缺失热图生成复原网络组成，在子网中保留 HRNet 原有的骨干网络，网络分支连接方式为并行连接，每个分支中的特征图分辨率从高到低依次排列，最后将多个不同尺度的特征图进行融合，重复利用特征图中的特征信息，在特征图后引入注意力机制提取关节点特征信息，网络最后的输出特征图传入复原网络模块进行特征信息的修复和补全，从而实现在复杂运动场景下的遮挡人像姿态估计。

1　高分辨特征生成复原网络

复杂运动场景的姿态估计任务需要克服的困难在于：1）运动员自身关节点遮挡问题，如图 1a 所示，运动员在运动锻炼过程中的动作比较复杂，自身的手臂关节或腿部关节点存在相互遮挡，以及关节点位置重叠，导致无法正确检测和定位被遮挡关节点；2）运动员之间的相互遮挡问题，如图 1b 所示，在多人竞技类体育运动和密集遮挡场所中，相互之间会遮挡一个或多个关节点，站位比较密集时的关节点分布也密集，易造成关节点错误连接；3）运动器械设备遮挡问题，如图 1c 所示，人与物的遮挡情况是关节点的部分缺失，需要根据人体骨架进行预测；4）复杂背景干扰问题，如图 1d 所示，当运动员的着装和背景颜色无法区分时会导致关节点的定位不准确或关节点缺失。

a. 自身遮挡　b. 相互遮挡　c. 物体遮挡　d. 背景干扰

图 1　复杂运动场景遮挡示例

1.1　网络模型结构设计

HRNet 网络在复杂遮挡场景下特征提取能力不足，遮挡关键点定位不够精确[19-20]，关节点遮挡对其他关节点的预测也产生负面影响[21]，从而导致 HRNet 网络在复杂遮挡场景下表现不佳。针对上述问题，新增融合注意力机制模块、多尺度特征图融合、反卷积和生成对抗模块，网络整体结构如图 2 所示。其中，HRNet 的主干骨架用于生成具有高分辨率图像信息的特征图，便于后续阶段的关节点预测和定位；多尺度特征图融合模块用于处理网络中的冗余信息和图像背景信息，完成大中小型人像和模糊目标人像的姿态估计；反卷积模块用于最大限度地复原特征信息；生成对抗模块的主要作用是在训练阶段增强热图的生成能力，通过学习真实样本的热图信息得到缺失部分的关节点分布，补全和预

测缺失的关节点信息，实现对处于复杂遮挡场景的人像关节点的精确预测。

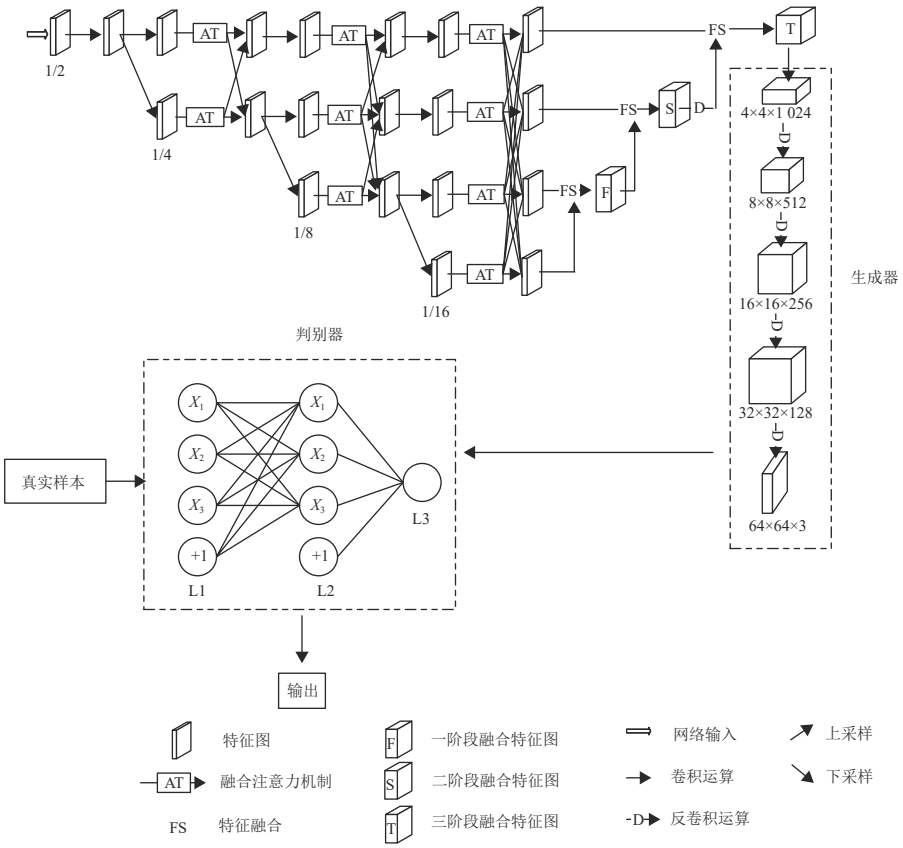


图 2 高分辨特征生成复原网络

主干骨架采用了并行连接方式，以保证各个阶段和各个分支都有不同分辨率的特征图。引入融合注意力机制，侧重关注特征图的人像区域和各个关节点的预测位置，避免冗余背景信息干扰造成关节点热图缺失。以多阶段融合的方式融合不同分辨率信息的特征图，得到信息完善的输出图像。高分辨特征生成复原网络的输出包含各类检测目标的图像信息和关节点位置信息，各个分支并行连接保证网络分支中存在各个分辨率的特征图，各个分支交互融合图像信息避免了背景干扰带来的错误预测结果。

1.2 融合注意力机制

融合注意力机制结合了通道注意力和空间注意力机制的优点，可以滤除图像背景信息的干扰，得到高准确率的关节点定位信息，融合方式如图 3 所示。

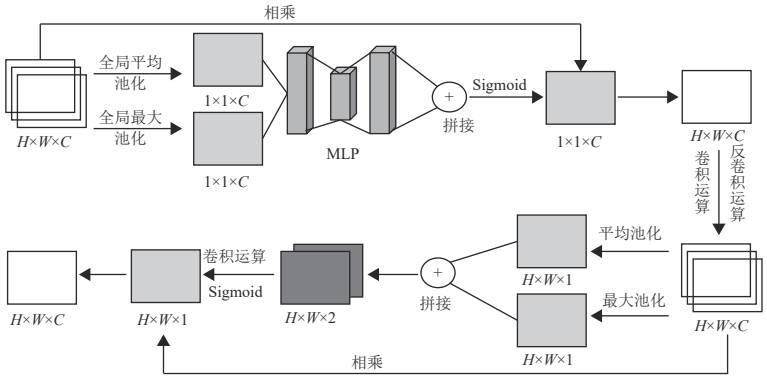


图 3 融合注意力机制示意图

具体实现方式为：尺寸为$H\times W\times C$的特征图分别做全局平均池化和全局最大池化，得到两部分$1\times 1\times C$特征图。随后将两部分$1\times 1\times C$特征图传入多层感知机 MLP 进行拼接，再用 Sigmoid 激活函

数层激活，得到$1\times1\times C$权重。最后将$1\times1\times C$权重与$H\times W\times C$特征图相乘，得到筛选出包含关键信息通道的特征图，大小为$H\times W\times C$。然后经过卷积运算和反卷积运算进一步滤除冗余信息。空间注意力机制模块分别对输入特征图进行最大池化和平均池化，得到两部分$H\times W\times1$特征图，然后将两部分$H\times W\times1$特征图拼接，得到$H\times W\times2$的特征图，通过卷积运算和 Sigmoid 激活得到$H\times W\times1$权重，再与$H\times W\times C$特征图相乘，以滤除背景信息，得到输出特征图，大小为$H\times W\times C$。经过 3 次重复操作，可以将包含关节点特征信息的通道和关键图像确定出来，得到进一步细化的特征信息，将遮挡缺失部分的关节点定位转化为对图像通道筛选和图像冗余信息去除的问题，简化了处理流程，减少了资源消耗。

1.3　多尺度特征图融合

经过融合注意力机制处理的特征图分辨率信息不一，各个分支的特征图中包含许多图像特征信息，需要进行整合后才能用于关节点定位与预测。因此，本文采用多尺度融合的方式得到最后阶段的特征图，用于关节点检测和定位。具体实现流程如图 4 所示。

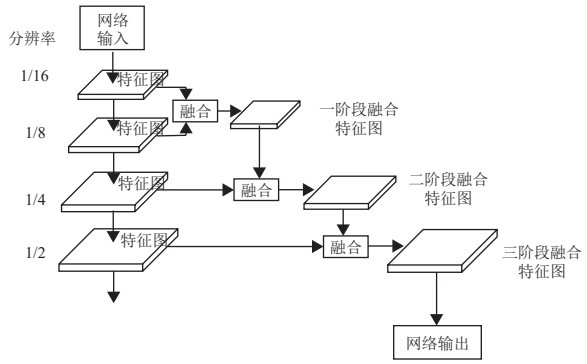


图 4　多尺度融合特征示意图

底层 1/16 分辨率的特征图，与 1/8 分辨率分支的特征图进行融合，得到一阶段的融合特征图；然后一阶段的融合特征图与 1/4 分辨率分支的特征图进行融合，得到二阶段融合特征图。二阶段的特征图信息可用于大型和中型目标人像的姿态估计，该阶段分辨率不高，包含的冗余信息较少，在网络训练时可以减少资源消耗；最后，二阶段特征图与 1/2 分辨率分支的特征图进行融合，得到三阶段融合特征图，该阶段特征图包含所有分辨率的信息，能够纠正前序预测中的错误信息，可用于小型目标和模糊目标的关节点检测与定位。网络整体的多尺度特征图融合能够充分利用图像信息，对于解决遮挡情况和复杂场景中的人体姿态估计起到推进作用。

1.4　生成对抗模块

本文根据生成对抗网络的思想，在基础网络上添加生成对抗模块，并在生成对抗模块中的生成器部分进行改进，添加反卷积运算来恢复特征图。改进后的生成对抗模块能够提升网络在热图缺失情况下的预测能力，根据数据集的关节点分布规律，学习人体关节点的热图分布。

生成对抗网络的目标函数为：

$$\min_G \max_D V(D,G) = E_{x\sim p_{\text{data}}(x)}[\log D(x)] + E_{z\sim p_{\text{noise}}(z)}[\log(1-D(G(z)))] \tag{1}$$

式中，G 表示生成器；D 表示判别器；$p_{\text{data}}(x)$代表真实样本分布；$p_{\text{noise}}(z)$代表低维噪声分布，通过生成器 G 映射到高维数据空间得到p_g。

固定生成器 G 的参数，优化判别器 D。即最大化 maxV(D, G)，等价于 min[$-V(D, G)$]，故判别器 D 的损失函数等价于：

$$J^{(D)}(\theta^D,\theta^G) = -E_{x\sim p_{\text{data}}(x)}[\log D(x)] - E_{\tilde{x}\sim p_g}[\log(1-D(\tilde{x}))] \tag{2}$$

固定判别器 D 参数时，生成器 G 的损失函数为：

$$J^{(G)} = \max_D V(G,D) = E_{x\sim p_{\text{data}}(x)}[\log D_G^*(x)] + E_{z\sim p_{\text{noise}}(z)}[\log(1-D_G^*(G(z)))] = E_{x\sim p_{\text{data}}(x)}[\log D_G^*(x)] + E_{\tilde{x}\sim p_g}[\log(1-D_G^*(x))] = E_{x\sim p_{\text{data}}(x)}\left[\log\frac{p_{\text{data}}(x)}{p_{\text{data}}(x)+p_g(x)}\right] + E_{x\sim p_g}\left[\log\frac{p_g(x)}{p_{\text{data}}(x)+p_g(x)}\right] \tag{3}$$

当$p_g = p_{\text{data}}$时，生成器的损失为：

$$J^{(G|D^*)} = E_{x\sim p_{\text{data}}(x)}\left[\log\frac{1}{2}\right] + E_{x\sim p_g}\left[\log\frac{1}{2}\right] = \log\frac{1}{2} + \log\frac{1}{2} = -\log 4 \tag{4}$$

引入 JS 散度，生成器的损失函数等价为：

$$J^{(G)} = E_{x\sim p_{\text{data}}(x)}\left[\log\frac{p_{\text{data}}(x)}{(p_{\text{data}}(x)+p_g(x))/2}\right] + E_{x\sim p_g}\left[\log\frac{p_g(x)}{(p_{\text{data}}(x)+p_g(x))/2}\right] - \log(4) = \text{KL}\left(p_{\text{data}}\|\frac{p_{\text{data}}+p_g}{2}\right) + \text{KL}\left(p_g\|\frac{p_{\text{data}}+p_g}{2}\right) - \log 4 = -\log 4 + 2\text{JS}D(p_{\text{data}}\|p_g) \tag{5}$$

由于 JS 散度具有非负性，所以当生成器样本与数据集样本分布相同时，散度为 0。$D(x)$ 的训练

效果与 $G(z)$ 的优化效果越好，生成器样本分布与真实样本分布越接近，损失就越小，判别器判别结果为真，否则判别为假。

生成对抗模块的具体结构如图 5 所示。在训练阶段，将白色掩膜噪声添加到输入图像形成遮挡来模拟关节点热图缺失。将处理后的图像传入生成器，生成器中包含 4 个反卷积块 DCONV，反卷积块用于最大限度还原高分辨率图像，4×4×1 024 尺寸的特征图逐级经过 DCONV 运算最后得到 64×64×3 的生成器输出特征图。判别器根据数据集样本的热图分布判断来自生成器的特征图是否为真，若判别结果为真，则输出生成器的特征图；若判别结果为假，则让生成器继续生成。

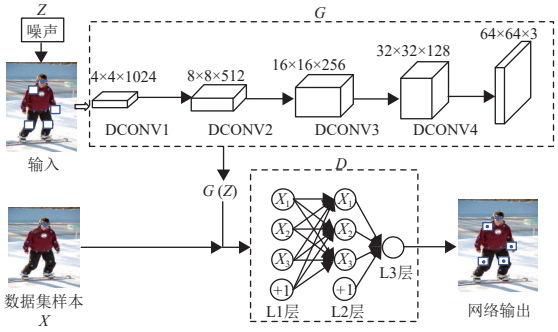


图 5 生成对抗模块

图 5 中以单幅图像为例，随机遮挡了人像的右肩、左腕、左膝和右膝 4 个关节点，特征图由 4×4×1 024 变为 64×64×3 传入到判别器。判别器根据学习到的热图分布规律进行判别，直至生成器的特征图符合真实样本的热图分布为止。最后，输出补全右肩、左腕、左膝和右膝 4 个关节点后的分布图。

2 实验结果与分析

2.1 实验数据集与评价指标

MSCOCO 数据集[22] 在人体姿态估计任务中标注出的人体关节点共计 17 个类别。数据集中共有 200 000 张图像，在这些图像中标记出共计 250 000 人的 17 个人体关节点。

Crowd Pose 数据集[23] 多用于解决拥挤场景的多人姿态估计问题，数据集中的标签人像被标注的关节点共计 14 个类别。数据集包含 20 000 张图片，涵盖 80 000 个被标注人像。

评价指标如下。

1）关键点相似性[24]（Object Key-point Similarity, OKS），计算真值和预测人体关节点的相似度：

$$\mathrm{OKS}_p=\frac{\sum_i \exp\{-d_{pi}^2/2S_p^2\sigma_i^2\}\delta(v_{pi}>0)}{\sum_i \delta(v_{pi}>0)} \tag{6}$$

式中，p 表示多人场景下图像中人物的编号；i 表示人体关键点的编号；d_{pi} 表示第 p 个人的第 i 个标注的关键点与模型预测的关键点结果之间的欧几里得距离；S_p 表示第 p 个人的尺度因子，数值为人体检测框面积的平方根；σ_i 表示第 i 个关键点的归一化因子，代表第 i 个关键点对整体的影响程度；v_{pi} 表示第 p 个人的第 i 个关键点的状态，共有 3 种状态，v_{pi}=2 时为可见，v_{pi}=0 时不可见，v_{pi}=1 时为遮挡；$\delta(\cdot)$ 是克罗内克函数，满足该函数条件要求时值为 1，不满足条件则值为 0。

2）正确关键点百分比[24]（Percentage of Correct Key-points, PCK），给定边界框内的候选区域包含原始关键点坐标位置，参照结合阈值得到的不同准确率来判断预测关节点的合理性，本文选择阈值 r=0.5。

3）平均精确度（Average Precision, AP）。

$$\mathrm{AP}^t=\frac{\sum_p \delta(\mathrm{OKS}_p>t)}{\sum_p 1} \tag{7}$$

式中，t 为给定 OKS 的阈值，本文选取阈值为 0.5 和 0.75；预测准确率由测试集所有图片中人物的 OKS 指标计算得到；当像素面积小于 32×32 时，以 AP^{S} 表示目标框所得的 AP 值；当像素面积介于 32×32～64×64 时，用 AP^{M} 表示 AP 的测量值；像素面积大于 96×96 的 AP 测量值则为 AP^{L}。

2.2 实验设置

实验使用 PyTorch 框架进行开发，GPU 选用 GTX 1080Ti，Epoch 设置为 200；训练数据集为 MSCOCO Dataset 和 Crowd Pose Dataset，输入图像大小设置为 256×192，Batch Size 为 32；优化器选用 Adam 优化器；learning rate=0.001；同时进行数据增强：预处理过程对图像进行随机角度旋转，角度范围选择为−45°～45°、图像随机尺度变换，变换范围为 0.65～1.35；随机翻转方向分为 X 轴、Y 轴、Z 轴 3 种；遮挡关节点数量在 0～25% 随机选取。所有网络都从随机初始化状态开始训练，没有经过预训练。

2.3 消融实验

在 HRNet 网络作为主干骨架的基础上设计了

消融实验，以验证各个网络模块的性能。针对 Crowd Pose 数据集，消融实验采用 32 通道宽度作为骨干网络，分析每个模块的精确度，实验结果见表 1。

表 1　各个网络模块的消融实验

反卷积	融合注意力机制	生成对抗模块	准确率/%
√			88.2
	√		88.4
		√	88.4
√	√		88.5
√		√	89.1
	√	√	89.6
√	√	√	**90.3**

所有模块都添加到骨干网络之后，反卷积模块前后的网络输出可以得到大中小型目标人像的有效信息，低分辨率的图像可用于预测大中小型目标人像，避免信息冗余，高分辨图像则用于小型目标和模糊目标的预测；融合注意力机制则能够对人像的关节点有所侧重，为遮挡图像的预测和缺失关节点的补全提供有效信息；生成对抗模块对于前序网络中的输出信息进行提取，缺失部分图像根据训练过程的学习来补全缺失部分的关节点做出高效正确的预测，测试图像准确率能达到 90.3%。

2.4　实验结果分析

为了验证本文网络的有效性，本文首先在 MSCOCO train 训练集上对高分辨特征生成复原网络进行训练，然后分别在 MSCOCO val 验证集和 MSCOCO test-dev 测试集上对改进网络与基线网络的性能进行测试。表 2 为本文网络与基线网络在 MSCOCO val 验证集上的性能比较结果，表 3 为本文网络与基线网络在 MSCOCO test-dev 测试集上的性能比较结果。其中，参数量表示网络模型的大小，GFLOPs 用于衡量网络的计算复杂度。

表 2　本文网络与基线网络在 MSCOCO val 验证集上的性能比较

网络模型	#Params/M	GFLOPs	输入尺寸	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}
HRNet-W32	28.5	7.1	256×192	65.9	86.4	70.6	66.5	57.9
HRNet-W48	**63.6**	14.6	256×192	66.8	86.7	71.1	66.7	58.4
Ours(HRNet-W32)	28.8	**48.1**	256×192	**69.9**	**90.8**	**76.5**	**66.9**	**76.9**

表 3　本文网络与基线网络在 MSCOCO test-dev 测试集上的性能比较

网络模型	#Params/M	GFLOPs	输入尺寸	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}
HRNet-W32	28.5	7.1	256×192	64.8	87.2	71.5	65.5	56.8
HRNet-W48	**63.6**	14.6	256×192	66.2	87.8	72.4	**66.0**	57.6
Ours(HRNet-W32)	28.8	**48.1**	256×192	**69.4**	**91.0**	**77.2**	65.7	**76.2**

根据表 2、表 3 中的结果可以看出，本文网络相较于基线网络 HRNet-W32 和通道数更多的 HRNet-W48，在模型参数量方面与 HRNet-W32 相近，小于 HRNet-W48，虽然计算量 GFLOPs 略大于基线网络 HRNet-W32 和 HRNet-W48，但在表 2 验证集上的 AP、AP^{50}、AP^{75}、AP^{M}、AP^{L} 均有不同程度的提升。同时在表 3 中可以看出，在测试集上 AP、AP^{50}、AP^{75}、AP^{L} 上均有提升，在 AP^{M} 指标上与 HRNet-W48 的最佳结果接近。

为了评估本文网络在复杂场景下的性能表现，本文在 Crowd Pose 数据集上对网络的性能进行了测试。首先，基于 Crowd Pose 训练集对卷积神经网络进行训练，然后，针对 Crowd Pose 测试集中不同场景下的图像进行姿态估计性能测试。Crowd Pose 数据集依据密集指数将图像中场景划分为稀疏、拥挤和密集 3 个水平，并在评估标准中使用 AP^{E}、AP^{M} 和 AP^{H} 来分别对各场景下的平均精确度进行表示。本文网络与基线网络在 Crowd Pose 测试集上的性能比较结果如表 4 所示。

根据表 4 结果可以看出，本文网络相较于基线网络 HRNet-W32 和通道数更多的 HRNet-W48，在模型参数量方面与 HRNet-W32 相近，小于 HRNet-W48，虽然计算量 GFLOPs 略大于基线网络 HRNet-W32 和 HRNet-W48，但是在 AP、AP^{50}、AP^{75}、AP^{M}、AP^{H} 上均有提升，在 AP^{E} 指标上与 HRNet-W48 的最佳结果接近。通过表 2、表 3 和表 4 实验结果可知，高分辨特征复原网络提高了对于关节点的估计精确度，在一定程度上可以有效克服复杂场景的干扰。

为了验证本文网络的先进性，分别在 MSCOCO

val 验证集和 MSCOCO test-dev 测试集上对本文网络与主流网络的性能进行测试，在 MSCOCO val 验证集和 MSCOCO test-dev 测试集上的性能比较结果分别如表 5、表 6 所示。

表 4　本文网络与基线网络在 Crowd Pose 测试集上的性能比较

网络模型	#Params/M	GFLOPs	输入尺寸	AP	AP^{50}	AP^{75}	AP^{E}	AP^{M}	AP^{H}
HRNet-W32	28.5	7.1	256×192	65.2	85.7	70.2	72.9	67.3	58.8
HRNet-W48	**63.6**	14.6	256×192	65.7	86.1	71.0	**73.7**	68.0	59.3
Ours(HRNet-W32)	28.8	**48.1**	256×192	**65.9**	**86.5**	**71.3**	73.5	**68.8**	**59.5**

表 5　本文网络与主流网络在 MSCOCO val 验证集上的性能比较

网络模型	输入尺寸	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}
HigherHRNet-W48	256×192	69.7	87.1	76.0	65.3	**77.0**
DEKR-W32	256×192	65.7	85.7	70.4	66.4	57.5
LiteHRNet-18	256×192	67.4	89.6	75.6	63.9	74.0
LiteHRNet-30	256×192	69.7	90.7	**77.5**	66.6	75.0
DiteHRNet-18	256×192	67.4	87.2	73.0	63.8	74.2
DiteHRNet-30	256×192	69.0	88.0	76.0	65.5	75.5
Ours(HRNet-W32)	256×192	**69.9**	**90.8**	77.3	**66.9**	76.9

表 6　本文网络与主流网络在 MSCOCO test-dev 测试集上的性能比较

网络模型	输入尺寸	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}
HigherHRNet-W48	256×192	68.8	87.9	76.3	62.5	76.1
DEKR-W32	256×192	65.0	86.4	70.6	65.5	57.0
LiteHRNet-18	256×192	66.5	89.9	74.4	62.7	73.1
LiteHRNet-30	256×192	67.6	88.7	76.9	64.8	74.4
DiteHRNet-18	256×192	67.1	88.0	73.5	61.9	73.5
DiteHRNet-30	256×192	68.3	89.4	**77.4**	64.9	75.2
Ours(HRNet-W32)	256×192	**69.4**	**91.0**	77.2	**65.7**	**76.2**

根据表 5 的结果可以看出本文网络相较于 HigherHRNet-W48、DEKR-W32、LiteHRNet-18、LiteHRNet-30、DiteHRNet-18、DiteHRNet-30，在 AP、AP^{50}、AP^{M} 上均有一定提升，在 AP^{75}、AP^{L} 两项指标上与其他网络的最佳结果十分接近。根据表 6 结果可以看出本文网络在 MSCOCO test-dev 测试集上相较于 HigherHRNet-W48、DEKR-W32、LiteHRNet-18、LiteHRNet-30、DiteHRNet-18、DiteHRNet-30 在 AP、AP^{50}、AP^{M}、AP^{L} 上均有一定提升，在 AP^{75} 指标上与其他网络的最佳结果十分接近。大型目标人像的关节点图像占比大，遮挡后导致图像整体信息丢失严重，不利于关节点的预测和估计，会影响 AP^{75} 和 AP^{L} 两项指标的变化；中小型目标人像的图像占比小，受遮挡处理影响之后的信息丢失较少，影响 AP、AP^{50}、AP^{M} 这 3 项指标的数值变化。而本文添加的注意力机制和生成对抗模块能够赋予网络预测和生成复原的能力，能够检测到其他网络检测不到的关节点，在遮挡情况下的关节点缺失问题也能依靠生成对抗模块进行复原，在中小型目标人像的姿态估计上体现出优越性。

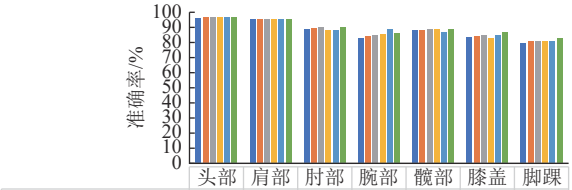
本文网络与主流网络在 Crowd Pose 测试集上的性能比较结果如表 7 所示。

表 7　本文网络与主流网络在 Crowd Pose 测试集上的性能比较

网络模型	输入尺寸	AP	AP^{50}	AP^{75}	AP^{E}	AP^{M}	AP^{H}
HigherHRNet-W48	256×192	65.6	86.4	70.6	73.3	68.1	58.9
DEKR-W32	256×192	65.7	85.7	70.4	73.0	66.4	57.5
LiteHRNet-18	256×192	64.4	83.0	69.8	72.8	67.2	58.5
LiteHRNet-30	256×192	65.7	85.8	70.3	73.1	68.5	58.7
DiteHRNet-18	256×192	64.6	84.8	69.0	71.2	67.6	55.9
DiteHRNet-30	256×192	65.8	86.0	70.5	73.2	68.7	58.8
Ours(HRNet-W32)	256×192	**65.9**	**86.5**	**71.3**	**73.5**	**68.8**	**59.5**

根据表 7 结果可以看出，本文网络在 MSCOCO test-dev 测试集上相较于 HigherHRNet-W48、DEKR-W32、LiteHRNet-18、LiteHRNet-30、DiteHRNet-18、DiteHRNet-30 在 AP、AP^{50}、AP^{75}、AP^{E}、AP^{M}、AP^{H} 上均有一定提升。能够对拥挤场景下的被遮挡人像进行准确姿态估计，对于缺失部分的关节点也能够通过网络训练得出补全的关节点位置，从而得到准确的估计结果，在解决拥挤场景和复杂场景的问题上效果较好。

为验证本文网络的预测效果，在 MSCOCO 数据集中的验证集图像进行关节点检测，各方法的预测精度对比如图 6 所示。其中，头部表示左眼、右眼、左耳、右耳、鼻 5 个人体关节点的平均准确率；肩部表示左肩和右肩两个关节点的平均准确率；肘部表示左肘和右肘的平均准确率；腕部表示左腕和右腕的平均准确率；髋部表示左髋和右髋的平均准确率；膝盖表示左膝和右膝的平均准确率；脚踝表示左脚踝和右脚踝的平均准确率。

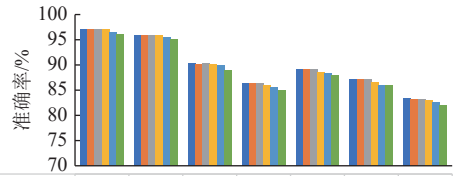


	头部	肩部	肘部	腕部	髋部	膝盖	脚踝
HRNet-W32	96.4	95.3	89.0	83.2	88.4	84.0	79.6
HigherHRNet-W48	96.9	95.9	89.5	84.4	88.4	84.5	81.3
DEKR-W32	97.0	95.9	90.0	85.0	89.2	85.3	81.3
LiteHRNet-30	97.1	95.8	88.4	86.0	88.9	83.0	81.0
DiteHRNet-30	97.1	95.8	88.0	88.9	86.9	84.9	81.2
本文(HRNet-W32)	97.1	95.9	90.3	86.4	89.1	87.1	83.3

图 6　关键点预测精度对比

由图中数据可以看出本文网络在头部、肩部、肘部、膝盖、脚踝关节点的预测准确率均高于 HRNet-W32、HigherHRNet-W48、DEKR-W32、LiteHRNet-30、DiteHRNet-30，对于腕部、髋部关节点的预测准确率分别达到 86.4% 和 89.1%。在普通场景下的人体姿态估计任务中，高分辨特征生成复原网络能够有效完成关节点的检测和姿态估计。

实验选取 Crowd Pose 数据集中的测试图像进行遮挡，遮挡关节点的数量按比例分别为 0、5%、10%、15%、20%、25% 进行测试，测试结果如图 7 所示。在几种遮挡率下头部关节点的检测精度能够保证在 96% 以上；对于图像面积占比较大的人体关节，如肩部、髋部和膝盖关节点的预测准确率在 85.9% 以上；在较小面积占比的人体部位如肘部、腕部、脚踝关节点的检测也均高于 82%；在不同遮挡率下的走势基本相同，不会出现因遮挡关节数量过多而导致无法检测的情况。这表明本文网络可以稳定地完成各种复杂拥挤场景下的姿态估计任务。



遮挡率/%	头部	肩部	肘部	腕部	髋部	膝盖	脚踝
0	97.1	95.9	90.3	86.4	89.1	87.1	83.3
5	97.1	95.9	90.2	86.3	89.1	87.1	83.2
10	97.1	95.9	90.3	86.3	89.1	87.1	83.1
15	97.1	95.9	90.2	85.9	88.6	86.5	83.0
20	96.5	95.5	89.9	85.6	88.4	86.0	82.5
25	96.0	95.0	89.0	85.0	88.0	85.9	82.0

图 7　不同遮挡率下关键点预测精度对比

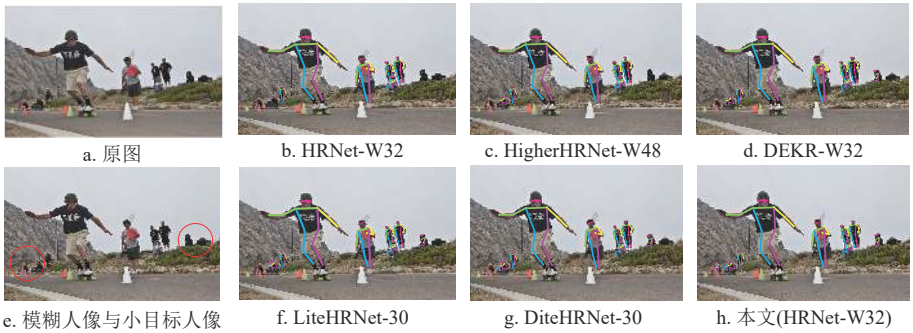
对本文网络在 Crowd Pose 验证集上的测试结果图做了可视化处理，单人姿态估计结果对比如图 8 所示。图中第一行为无遮挡情况的单人姿态估计结果，本文网络和 DiteHRNet-30 能够检测到人像的左膝和右膝两个关节点并进行正确连接，其余网络则漏检了这两个关节点。图中第二行为有遮挡情况的单人姿态估计结果，由于人像侧身站位，遮挡住了左肘关节点，各个网络都能够检测出人像的左腕关节点，本文网络在检测出人像的左腕关节点的同时，还能够基于网络的预测能力预测出左肘关节点，完整连接出人像的左手骨架。

图中第三行为较复杂遮挡情况的单人姿态估计结果，HRNet-W32 错误估计了左踝和右踝的位置，在连接时也将左膝和右膝跟后续关节连接错误；其他网络虽能正确估计左踝和右踝的位置，但是左膝的位置预测与实际存在偏差；本文网络能够正确估计左踝、右踝和左膝的位置，在关节点预测和骨架连接的准确性上都优于其他网络。

由图 9 中的原图预测结果可以看出，当所拍摄图像中的人像大小不一，且目标人像较小和较模糊时，其他网络对于原图中最左边和最右边的小目标人像无法预测完全。而本文网络不仅可以检测和估计原图中左边的两个小目标人像，还可以预测和估计原图中最右边的小目标人像，对于小目标和模糊目标预测表现均优于其他几个网络。

a. 原图 b. HRNet-W32 c. HigherHRNet-W48 d. DEKR-W32 e. LiteHRNet-30 f. DiteHRNet-30 g. 本文(HRNet-W32)

图 8 单人姿态估计结果对比



a. 原图 b. HRNet-W32 c. HigherHRNet-W48 d. DEKR-W32

e. 模糊人像与小目标人像 f. LiteHRNet-30 g. DiteHRNet-30 h. 本文(HRNet-W32)

图 9 模糊目标与小目标人像的姿态估计结果对比

图 10 是在复杂运动场景下的可视化结果，以 Crowd Pose 数据集为基准预测人体的 14 个关节点位置并可视化到输出结果图中。由可视化结果可以看出本文网络能够将处于遮挡情况和复杂背景情况的关节点找出并正确连接出人体姿态骨架。从预测结果看来，本文网络能够处理室内场景和室外场景的人体姿态估计，也能够有效捕捉小型目标人像和模糊人像的特征信息，把握关节点信息并精准定位。

图 10 运动场景下的多人姿态估计

3 结束语

本文以 HRNet 网络作为基础网络，加入反卷积模块、融合注意力机制和生成对抗模块，将前序的高分辨特征图网络输出作为后序的缺失热图生成复原网络的输入，最终的输出结果作为姿态估计的网络输出，整体的网络为高分辨特征生成复原网络。使用 MSCOCO Dataset 和 Crowd Pose Dataset 进行网络训练和测试，所得的实验结果表明：提出的姿态估计方法在普通场景下准确率高，对不同的人体关节点预测精度均有不同程度的提升，对于 0～25% 程度的关节点遮挡及复杂运动场景模拟情况，具有优越稳定的效果。未来的研究方向是保证准确率的同时，做到网络的轻量化。

参 考 文 献

[1] 胡青松, 张亮, 丁娟, 等. 人体动作数据编码与 CNN 精确识别[J]. 电子科技大学学报, 2020, 49(3): 473-480.
HU Q S, ZHANG L, DING J, et al. Data encoding and CNN accurate recognition of human body motion[J].

Journal of University of Electronic Science and Technology of China, 2020, 49(3): 473-480.

[2] ZHANG J H, LI P, JIN C C, et al. A novel adaptive Kalman filtering approach to human motion tracking with magnetic-inertial sensors[J]. IEEE Transactions on Industrial Electronics, 2020, 67(10): 8659-8669.

[3] BISSACCO A, YANG M H, SOATTO S. Fast human pose estimation using appearance and motion via multi-dimensional boosting regression[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2007: 1-8.

[4] FISCHLER M A, ELSCHLAGER R A. The representation and matching of pictorial structures[J]. IEEE Transactions on Computers, 1973, C-22(1): 67-92.

[5] TOSHEV A, SZEGEDY C. DeepPose: Human pose estimation via deep neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2014: 1653-1660.

[6] WEI S H, RAMAKRISHNA V, KANADE T, et al. Convolutional pose machines[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2016: 4724-4732.

[7] RAMAKRISHNA V, MUNOZ D, HEBERT M, et al. Pose machines: Articulated pose estimation via inference machines[M]//Lecture Notes in Computer Science. Cham: Springer International Publishing, 2014: 33-47.

[8] NEWELL A, YANG K Y, DENG J. Stacked hourglass networks for human pose estimation[M]. Cham: Springer International Publishing, 2016: 483-499.

[9] CHU X, YANG W, OUYANG W, et al. Multi-context attention for human pose estimation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2017: 5669-5678.

[10] KREISS S, BERTONI L, ALAHI A. PifPaf: Composite fields for human pose estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2019: 11977-11986.

[11] 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. New York: IEEE, 2019: 5693-5703.

[12] CHENG B W, XIAO B, WANG J D, et al. Bottom up higher resolution networks for multi-person pose estimation[EB/OL]. [2023-03-27]. https://doi.org/10.48550/arXiv.1908.10357.

[13] 吴春梅, 胡军浩, 尹江华. 利用改进生成对抗网络进行人体姿态识别[J]. 计算机工程与应用, 2020, 56(8): 96-103.
WU C M, HU J H, YIN J H. Using improved generative adversarial network for human pose estimation[J]. Computer Engineering and Applications, 2020, 56(8): 96-103.

[14] YU C Q, XIAO B, GAO C X, et al. Lite-HRNet: A lightweight high-resoultion network[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 10435-10445.

[15] GENG Z G, SUN K, XIAO B, et al. Bottom-up human pose estimation via disentangled keypoint regression [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2021: 14671-14681.

[16] LI Q, ZHANG Z Y, XIAO F, et al. Dite-HRNet: Dynamic lightweight high-resolution network[EB/OL]. [2023-03-27]. https://doi.org/10.48550/arXiv.2204.10762.

[17] CHEN H Y, HE J Y, XIANG W M, et al. HDFormer: High-order directed transformer for 3D human pose estimation[EB/OL]. [2023-03-27]. https://doi.org/10.48550/arXiv.2302.01825.

[18] YANG J, ZENG A L LIU S L, et al. Explicit box detection unifies end-to-end multi-person pose estimation [EB/OL]. [2023-03-27]. https://doi.org/10.48550/arXiv.2302.01593.

[19] 毋宁. 基于改进高分辨率网络的人体姿态估计算法研究[D]. 西安: 西安工业大学, 2023.
WU N. Research on human posture estimation algorithm based on improved high resolution network[D]. Xi'an: Xi'an Technological University, 2023.

[20] 宾言锐. 基于深度学习的人体姿态估计方法[D]. 武汉: 华中科技大学, 2020.
BIN Y R. Human pose estimation method based on deep learning[D]. Wuhan: Huazhong University of Science and Technology, 2020.

[21] 褚真, 米庆, 马伟, 等. 部位级遮挡感知的人体姿态估计[J]. 计算机研究与发展, 2022, 59(12): 2760-2769.
CHU Z, MI Q, MA W, et al. Part-level occlusion-aware human pose estimation[J]. Journal of Computer Research and Development, 2022, 59(12): 2760-2769.

[22] LIN T Y, MICHAEL M, et al. Microsoft coco common objects in context[DB/OL]. [2023-03-27]. https://d-oi.org/10.48550/arXiv.1405.0312.

[23] LI J F, WANG C, ZHU H, et al. Crowd pose: Efficient crowded scenes pose estimation and a new benchmark[EB/OL]. [2023-03-27]. https://doi.org/10.48550/arXiv.1812.00324.

[24] YANG Y, RAMANAN D. Articulated pose estimation with flexible mixtures-of-parts[C]//Proceedings of the CVPR 2011. New York: IEEE, 2011: 1385-1392.

编　辑　叶　芳