文章编号:1671-4598(2025)02-0238-08 DOI:10.16526/j.cnki.11-4762/tp.2025.02.030 中图分类号:TP391.41 文献标识码:A

基于 PCNet 的人体姿态估计方法

马洋平[1], 曹 薇[2], 展宗思[2], 徐志君[1], 王有发[3]
(1. 浙江浙大网新众合轨道交通工程有限公司, 杭州 310051;
2. 西安市轨道交通集团有限公司, 西安 710016;
3. 西安交通大学 软件学院, 西安 710049)

摘要: 人体姿态估计是计算机视觉、模式识别领域的重要研究问题, 用于将视频图像中的人体骨骼姿态进行检测识别, 在人机交互领域有重要应用; 针对闸机场景下人群拥挤、遮挡严重的人体姿态估计问题, 提出了基于姿态矫正的人体姿态估计网络 PCNet; 该网络设计了一种融合全局和局部信息的 Transformer 特征编码模块, 并将其引入到模型特征提取骨干网络中提升精度表现; 提出基于时空注意力机制的级联结构的姿态矫正模块, 对预测的关键点位置进行矫正, 修正因遮挡、小尺度目标等引起的误差较大的关键点; 将提出的人体姿态估计方法在 COCO 数据集和 CrowdPose 数据集上进行实验, 实验结果表示, 模型效果与主流方法相比在精度和鲁棒性上均得到了提升。

关键词: 人体姿态估计; Transformer; 复杂场景; 姿态矫正; HRNet

Human Pose Estimation Based on PCNet

MA Yangping[1], CAO Wei[2], ZHAN Zongsi[2], XU Zhijun[1], WANG Youfa[3]
(1. Zhejiang Zheda Insigma Rail Transportation Engineering Co., Ltd., Hangzhou 310051, China;
2. Xi'an Rail Transit Group Co., Ltd., Xi'an 710016, China;
3. School of Software Engineering, Xi'an Jiaotong University, Xi'an 710049, China)

Abstract: Human pose estimation is an important research problem in the field of computer vision and pattern recognition, which is used to detect and recognize human skeleton posture in video images, and has important applications in the field of human-computer interaction. Aiming at the human pose estimation of crowd congestion and severe occlusion in gate scenarios, a human pose estimation network PCNet based on posture correction is proposed. In this network, a transformer feature coding module that combines global and local information is designed, which is introduced into the backbone network of model feature extraction to improve the accuracy. The proposed attitude correction module with cascade structure based on spatio-temporal attention mechanism corrects the predicted key points, and adjusts the key points with large errors caused by occlusion and small-scale targets. The proposed human pose estimation method is tested on COCO dataset and CrowdPose dataset. Experimental results show that the method improves the accuracy and robustness of the model compared with the mainstream methods.

Keywords: human pose estimation; Transformer ; complex scenes; pose correction; HRNet

0 引言

在社会生活中, 姿态和行为是人与人之间除语言之外沟通和传递信息、表达情感的重要方式。通过对人类的姿态进行识别和分析, 对于揭示个体的行为动机和心理状态至关重要。在这样的背景下, 人体姿态估计(HPE, human pose estimation) 是计算机视觉研究中的一个重要领域, 它在动作识别、人机交互及安全监控等方面展现出了重要的应用潜力[1-2]。

人体姿态估计是从 RGB 图像中检测出人体骨骼关键点坐标, 并将骨骼关键点按人体骨架模型顺序连接构成人体姿态的任务。2012 年, AlexNet 在 ImageNet 分类任务中取得突破, 推动了深度学习的复兴[3]。随后, DeepPose 利用深度学习技术显著提高了人体姿态估计的准确性, 为该领域带来了新的发展机会[4]。如今, 姿态估计研究主要基于简单场景, 而生活中除了日常简单场景之外, 同时也有车站、商场、地铁等人流量大、人群密集且遮挡严重的复杂环境, 因此, 准确捕捉和描述

收稿日期:2024-08-29; **修回日期**:2024-10-18。
作者简介:马洋平(1988-),男,大学本科,工程师。
引用格式:马洋平,曹 薇,展宗思,等.基于 PCNet 的人体姿态估计方法[J].计算机测量与控制,2025,33(2):238-245.

人体姿态是该领域面临的重要挑战[5]。然而当然着眼于遮挡问题的方法不多且精度不高，所以针对这类复杂场景的人体姿态估计任务仍然有较大发展空间。多人姿态估计根据关键点检测思路的不同，可分为自底向上（Bottom-up）和自顶向下（Top-down）两类[6-8]。

自顶向下方法：先检测个体，再对每个个体进行关键点定位，精度高但速度慢。例如 G-RMI、RMPE、Coarse-Fine Network[9-14]等尝试通过改进网络结构来提高精度，但未充分考虑关键点检测阶段的遮挡。这些方法虽然针对遮挡问题做出优化但是检测精度不够高。

自底向上方法：直接预测图像中所有人的关键点，然后将它们关联到各自的人体。速度快，但平均精度较低。如 DeepCut、OpenPose、PifPaf 及 SAHR 等方法[15-18]，这些方法主要关注提高自底向上方法的精度，并没有针对遮挡问题。

在多人二维姿态估计中，还有一些方法既不属于传统的自顶向下，也不属于自底向上的策略。如 DLCM 改善部件间联系[19]，SPM 简化流程[20]，以及利用 Transformer 捕捉全局依赖性的方法[21-23]。上述方法往往针对网络本身的速度和精度的平衡以及 Transformer 结构在姿态估计任务中的使用探索，对于方法的应用场景并没有特别关注。

为了解决现有姿态估计算法在人群密集、遮挡严重的场景下出现的问题，本文设计了一种融合全局和局部信息的 Transformer 结构特征编码模块，将其引入到姿态估计模型的特征提取骨干网络中，更好地利用上下文信息；此外，提出了一种基于时空注意力机制和级联结构的姿态矫正方法，对复杂场景中预测误差较大的关键点进行进一步修正，从而得到一种针对人群拥挤、肢体遮挡的复杂场景的高精度人体姿态估计方法。

1　基于姿态矫正的人体姿态估计算法

1.1　PCNet 的整体结构

研究的人体姿态估计任务主要针对日常生活中的复杂场景，例如地铁站等人流量较大的场所，该场景下的人体目标常常比较密集且尺度大小不一，同时人体之间也经常出现自遮挡和互遮挡等情况，遮挡情况如图 1 所示。目的是提出一种高度适应于复杂场景，同时具有高泛化能力，可以更好地适应于各种真实场景的人体姿态估计网络。基于以上提出的问题和应用场景，本论文设计了一种两阶段的、自顶向下的人体姿态估计网络，其算法流程如图 2 所示。

方法对输入的 RGB 图像使用 Faster RCNN 目标检测网络对所有人体目标进行检测，检测完成后根据目标检测框裁减出图像中的每个人体目标，之后通过提出的两阶段人体姿态估计网络 PCNet 对于目标检测器检测到的每个人体目标通过基于热图回归的单人姿态估计方法预测其人体关键点的坐标位置，并将各个关键点根据人体骨架模型两两相连，组合为人体骨架模型。

图 1　现有方法在人群拥挤、遮挡场景的错误案例

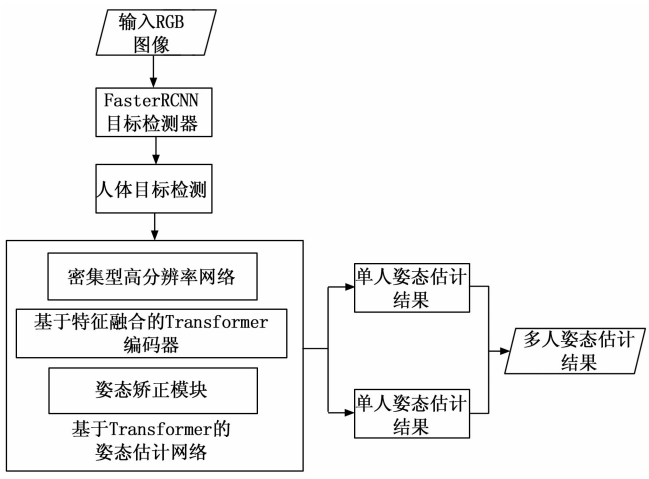


图 2　人体姿态估计流程图

单人姿态估计网络部分的网络结构如图 3 所示，针对浅层的特征提取网络，采用了卷积神经网络（CNN，convolutional neural network），其结构类似高分辨率人体姿态估计网络 HRNet[13]，在此基础上改进了高分辨率人体姿态估计网络的结构，将相同尺寸的特征图间的串联方式更改为密集连接。并且设计了一种融合全局和局部信息的 Transformer 特征编码模块来更好地进行特征提取，除此之外，使用改进的密集型高分辨率网络对人体关键集完成初始阶段的预测后，一种基于时空注意力的级联结构的姿态矫正模块为复杂场景下的自遮挡、互遮挡等困难关键点进一步矫正，有效地提高此类关键点预测的精确度，对于复杂场景有更好的效果和适应性。

人体姿态估计旨在从输入图像中检测 K 个人体关键点的位置，现有大多数方法利用 K 个热力图来表示这些关键点，其中热力图的像素点值表示当前空间位置存在人体关键点的概率。

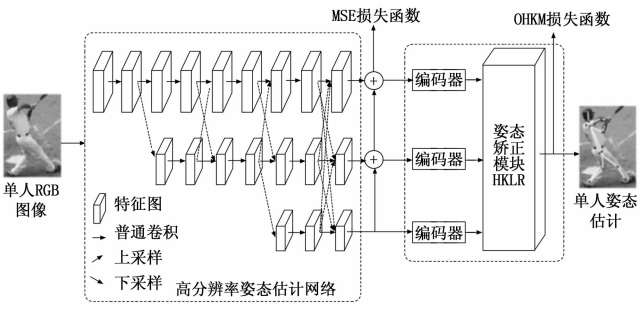


图 3 基于姿态矫正的人体姿态估计网络结构

$$X_j = FastRcnn(Input) \tag{1}$$

$$F = HRNet(X_j) \tag{2}$$

式中，$X_j \in R^{h\times w\times 3}$ 表示一个人体检测框的图像，$Input \in R^{H\times W\times 3}$ 表示输入的图像，$FastRcnn$ 表示目标检测器。$F \in R^{h\times w\times K\times I}$ 表示初步姿态估计的结果，K 为每个实例骨骼点的个数，本文中为 17；F_i 表示一个人体姿态，其中 $i \in 1, 2, \cdots, I$，I 表示 HRNet 网络分支的个数，每个分支都会得到一个人体姿态。

$$P_i = HRKL[E(F_i)] \tag{3}$$

$$Y = FC[concate(P_1, P_2, \cdots, P_I)] \tag{4}$$

式中，$F_i \in R^{h\times w\times K}$ 表示 HRNet 一个分支的结果，E 表示基于特征融合的 Transformer 编码器，$HRKL$ 表示基于时空注意力机制的姿态矫正模块。P_i 表示第 i 个分支姿态纠正的结果；$concate$ 表示张量拼接操作；FC 表示全连接层；$Y \in R^{h\times w\times K}$ 表示预测出的人体姿态。

1.2 基于特征融合的 Transformer 编码器

全局信息是指图像或特征图的整体结构和上下文信息。这包括人体在整个场景中的位置、姿态以及与其他物体的关系。例如，在地铁站这样的人群密集场景中，全局信息可以帮助模型理解每个人体目标之间的相对位置，从而更好地处理遮挡问题。局部信息则关注于特定区域内的细节和纹理。它涉及单个关键点及其周围环境的精细特征。例如，对于一个被部分遮挡的手臂，局部信息有助于准确识别手臂的关键点，即使这些关键点在整体上并不显眼。

现有研究多人姿态估计的相关工作中，卷积神经网络是最为常见的特征提取网络，但卷积缺乏捕获图像全局信息的能力，无法建模特征之间的依赖和联系，从而不能充分地利用上下文信息。相比之下，Transformer 的自注意力机制能够挖掘长距离的依赖关系，实现全局关系建模[24]。近年来 Transformer 模型在许多视觉任务中展现出巨大潜力，受这些成果的启发，本章设计了一种融合全局和局部信息的 Transformer 结构特征编码模块，引入到姿态估计模型的特征提取骨干网络中，提升了模型在人群拥挤、遮挡的复杂场景下的精度表现。本章提出的融合全局和局部信息的 Transformer 特征编码模块计算公式如下所示：

$$F'_i = F_i + FLA[LayerNorm(F_i)] \tag{5}$$

$$E(F_i) = F'_i + FFN[LayerNorm(F'_i)] \tag{6}$$

式中，$F_i \in R^{h\times w\times K}$ 表示输入特征；$F'_i \in R^{h\times w\times K}$ 表示中间特征；$E(F_i)$ 表示输出特征；FLA 表示全局和局部自注意力模块；$LayerNorm$ 表示层归一化；FFN 表示前向传播网络特征编码模块的网络结构如图 4 所示，该模块包含全局和局部自注意力模块（FLA），层归一化（LayerNorm）和前向传播网络（FPN，forward propagation network）。首先，输入特征经过层标准化操作进行标准化处理，然后被送入全局和局部自注意力模块，全局和局部自注意力模块会与输入特征进行相加操作得到中间特征。最后，中间特征通过带有残差连接和层归一化的前向传播网络进行处理，得到最终的 Transformer 编码器模块的输出结果。该模块保持输入与输出特征的空间维度和通道维度不变。其中前向传播网络包括两层全连接层、两层 Dropout 层以及 GELU 激活函数。

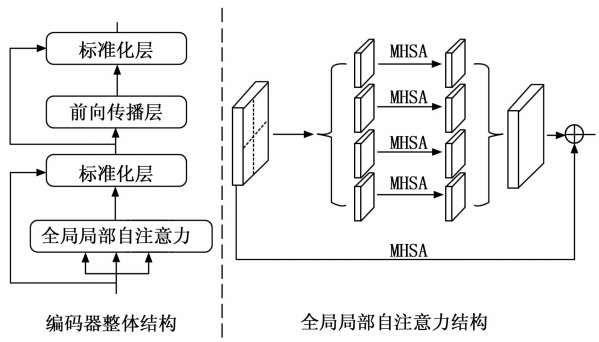


图 4 基于特征融合的 Transformer 编码器

全局和局部自注意力模块主要由两个分支组成，如图 4 所示，上层分支为局部注意力分支，将输入的特征图划分为 4 个不重叠的小窗口，分别对每窗口区域的特征施加多头注意力（MHSA，multi-head self attention）操作[24]，然后再将输出特征按序拼接得到完整的输出。下层分支为全局注意力分支，将输入特征直接送入多头注意力模块进行全局信息提取，最后使用相加的方式融合局部信息特征和全局信息特征。

1.3 基于空间注意力机制的姿态矫正模块

输入的 RGB 图像经过上述网络后得到初始预测的人体关键点位置。在姿态矫正阶段，针对图像中的遮挡困难点提出了一种基于时空注意力机制和级联结构的姿态矫正方法，对复杂场景中预测误差较大的关键点进行进一步修正，从而达到在自遮挡或互遮挡的情况下，关键点的估计依然可以保持较高的精度。姿态矫正采用级联结构，针对不同分辨率的特征图，经过多个基于空间注意力的瓶颈块（Bottleneck Block），不同分辨率的特

征图通过不同倍率的上采样，对各个分辨率间的特征图进行连接和融合，其结构如图 5 所示。

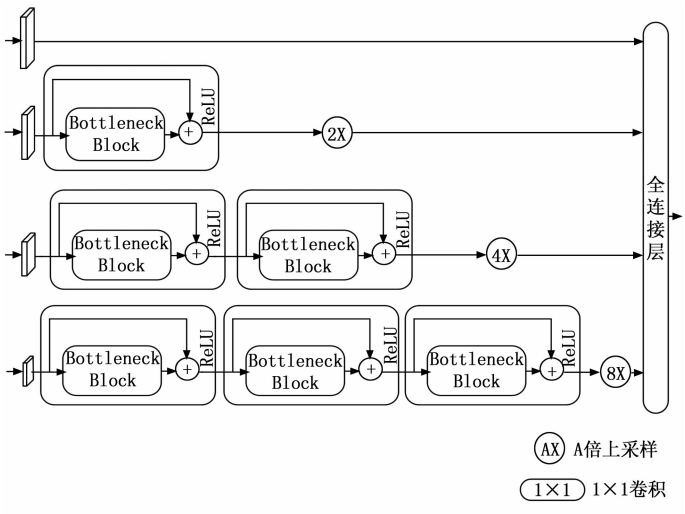


图 5　基于空间注意力机制的姿态矫正模块

采用级联的形式将所有分辨率的特征图连接起来，使得更深的网络层次能够处理更小尺寸的空间特征图，从而在提高计算效率的同时更好地平衡模型的有效性和复杂性。瓶颈块的具体结构如图 6 所示，其中利用多头自注意力模块来提取空间特征，首先将输入的特征图重排为二维矩阵形式，其中每一行对应于特定关键点的特征向量。通过线性变换，从这一特征矩阵中生成查询（Query）、键（Key）和值（Value）三类矩阵，并进一步分割成多个独立的头。每个头执行自身的注意力计算，从而允许模型并行地捕捉不同尺度的空间信息。例如，在分析跑步动作时，某些头部可能专注于局部细节如手部或脚部的位置，而其他头部则侧重于整体结构，如躯干方向与腿部弯曲程度。

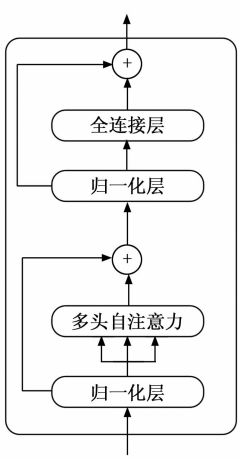


图 6　瓶颈块结构

这种多头设计促进了对复杂依赖关系的学习，不仅能够准确识别关节位置，还增强了处理遮挡情况的能力。通过加权各位置特征，模型能够赋予关键区域更高的关注度，即使存在背景干扰或部分身体被遮挡，也能利用上下文信息进行精确的姿态推断。例如在一个人的左肩被遮挡时，模型可以通过右肩或其他相关部位的信息来推测左肩的位置。最终，来自各个头的输出被聚合并通过一个额外的线性层转换回原始维度，产生具有增强空间特性的特征图。公式如下：

$$HRKL(F_i) = upsample\{B\cdots B[(F_i)]\} \tag{7}$$

式中，$F_i \in R^{h\times w\times d}$ 表示输入特征；$F'_i \in R^{L\times d}$ 表示中间特征；$B(F_i)$ 表示瓶颈块；$upsample$ 表示上采样。

1.4　在线困难点挖掘算法的均方误差损失函数

在网络的初始预测阶段，采用类似 HRNet[13] 的处理方式，使用 MSE 损失函数[25]来衡量预测热图与实际热图之间的差异。其中高斯核的标准差为 1 像素。

随着训练深度的增加，网络往往会更加关注于大多数简单的、预测难度较低的关键点，而对于其中的困难关键点的关注度则逐渐降低，例如图像中自遮挡或互遮挡区域的关键点等等。所以在姿态矫正阶段，需要对困难点有效地进行选择并通过反向梯度传播进一步定位。因此，参考在线困难样本挖掘（OKEM，online hard example mining）算法，选择使用基于在线困难点挖掘（OHKM，online hard keypoint mining）算法的损失函数，来对均方误差损失进行排序，筛选出最困难的前 8 个关键点进行重点回归。MSE 损失函数公式如下：

$$J_{\mathrm{MSE}} = \frac{1}{N}\sum_{i=1}^{N} -(y_i - \hat{y}_i)^2 \tag{8}$$

$$J_{\mathrm{OKHM}} = \frac{1}{N}\sum_{i=1}^{N} w\cdot(m_i - \hat{m}_i)^2 \tag{9}$$

其中：N 代表一个批次中样本个数，y_i 代表第 i 个样本的预测姿态；$\hat{y}_i$ 代表第 i 个样本的真实姿态；m_i 代表与真值差距最大的八个关键点的位置；w 代表困难点的损失权重。

2　实验结果与分析

2.1　人体姿态估计数据集

本章所提出的人体姿态估计方法的实验主要分为两部分，第一部分的实验针对日常场景下的人体姿态估计，这一部分是基于微软公司提供的 COCO 数据集[26]进行实验，使用的是该数据集 2017 年扩充的版本。第二部分的实验是针对人群密集，遮挡严重的复杂场景下的人体姿态估计，这部分实验在复杂场景下的人体姿态估计任务中应用较为广泛的 CrowdPose 数据集[27]完成。

2.1.1　COCO 数据集

COCO 数据集的图像主要选自于日常生活的场景，包含 91 个类别，共计约 12 万张图像，其中分别为 11.5 万训练集图像和 5 千张验证集图像。COCO 数据集共有目标实例、关键点、看图说话 3 种标注类型，本章的实

验主要基于目标关键点的标注数据进行训练。在 COCO 数据集中人体关键点的数量为 17 个，其具体格式如图 7（a）所示。

2.1.2 CrowdPose 数据集

为了有效地验证提出的模型在人群密集、遮挡严重的复杂场景下的人体姿态估计能力，第二部分的实验基于 CrowdPose 数据集完成。CrowdPose 数据集共包含 2 万张图片，8 万个行人，训练、验证和测试集按照 5∶1∶4 的比例进行划分。CrowdPose 数据集提出了使用密集度（CI，crowd index）来表示图像中人群的密集程度，其计算如公式（10）所示：

$$Crowd\ Index = \frac{1}{n}\sum_{i=1}^{n}\frac{N_i}{N_i} \tag{10}$$

其中：i 是第 i 个人体目标的目标检测框；$\frac{N_i}{N_i}$ 分别表示属于该人体目标的关键点个数和不属于该人体目标的关键点个数；$\frac{N_i}{N_i}$ 第 i 个行人实例的 CI 值。CrowdPose 数据集的组建是根据 CI 将公开数据集的图像分为 20 组，从 0 到 1，不同组的 CI 步长为 0.05，从中均匀采样 2 万张图像。CrowdPose 数据集标注人体关键点数量为 14 个，具体格式如图 7（b）所示。

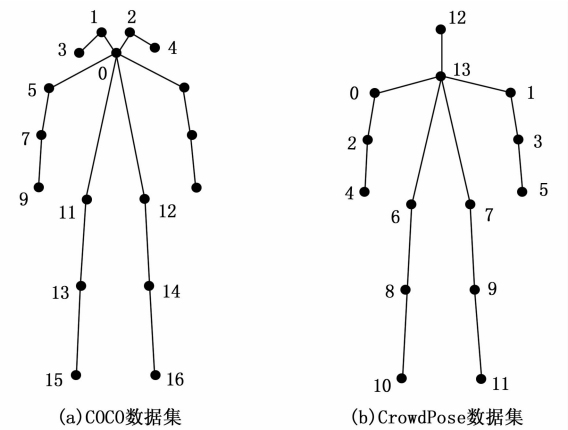


图 7 人体关键点标注格式

CrowdPose 数据集的图像既能覆盖密集人群的场景也可以覆盖日常场景，CrwodPose 数据集的平均 IoU 为 0.27，而 MSCOCO 数据集、MPII 数据集[28]、AIChallenger 数据集[29]的 IoU 只有 0.06、0.11 和 0.12，这也可以证明 CrowdPose 数据集的图像的人体目标密集程度要远远高于其他 3 个数据集，更符合针对复杂场景的目标，可以充分地验证提出的模型在复杂场景下的有效性。

2.2 评价指标

平均准确率（AP，average precision）表示精度百分比，表示对象关键点相似度 OKS 大于阈值 T 的关键点个数占所有预测关键点的比例。针对自顶向下的 2D 多人姿态估计，其过程是先找出图像中的所有人体目标，之后针对每个目标进行关键点的检测，具体如式（2）所示：

$$AP = \frac{\sum_p r\delta(OKS_{p>T})}{\sum_p r1} \tag{11}$$

其中：OKS_p 表示第 p 个关键点的 OKS 值，T 为人工标定的阈值。

平均准确率均值（mAP，mean average precision）是常用检测指标，通过给 AP 指标设置不同的阈值 T，例如\mathrm{T∈}⊢[⊣r{0.5:0.05:0.95}⊣]，就可以获得多个 AP 指标，通过对所有 AP 指标求均值获得 mAP 值。

AP 和 AR 指标是人体姿态估计常用的评价指标，本文实验中的 AP 指标代表平均准确率均值 mAP 指标，AP∶50 和 AP∶75 表示对象关键点相似度 OKS 阈值设定为 0.50 和 0.75 时的平均准确率 AP 指标。

2.3 实验结果比较和分析

由于本章提出的人体姿态估计模型的应用场景为人群密集、遮挡严重的复杂场景，因此本章实验主要基于包含日常场景的 COCO 数据集和包含的复杂场景的 CrowdPose 数据集，实验从定性和定量两个部分展开，基于 COCO 数据集完成对比实验，基于 CrowdPose 数据集完成对比实验和消融实验，并采用不同阈值的人体关键点平均准确率 AP 和平均召回率 AR 指标对模型的姿态估计效果进行分析和评估。充分验证本文提出的模型在日常场景和复杂场景两个应用场景下的有效性。本章实验的运行环境为 Ubuntu Linux 系统，硬件配置为 24 G 显存的英伟达 GeForce RTX 3090 图形处理器和 2.9 GHz 的 Inter（R）Xeon（R）Gold 6226R 中央处理器。

首先使用 COCO 数据集进行对比实验，将人体目标检测框大小设置为纵横比 4∶3，并将裁减出的检测框调整至尺寸为 384×288 的图像作为输入，使用 Adam 优化器，超参数设置如下：batch 大小为 16，epoch 数目为 210，初始学习率大小为 1×10^{-4}，并且分别在第 170 和第 200 个 epoch 时将学习率下降至 1×10^{-4} 和 1×10^{-5}。实验结果如表 1 所示。

表格中的（一）代表未知的数据。由表 1 结果可以看出，本论文提出的方法对比不同的人体姿态估计方法均取得了最好的效果。相比于 HRNet 基准网络平均准确度均值 AP 大概提高了 2.2%，相比于多实例预测网络 MIPNet，PCNet 平均精度提高了 0.9%，相比于近年提出的 HR-Former-B 等 Transformer 结构的方法，TMI-Net 在平均精度 上提高了约 2.7%。相比于最新的自顶向下算法和自底向上算法，例如 AD-HNN、DPIT、

表 1　COCO 数据集上的实验结果

算法	AP	AP^{50}	AP^{75}	AR
HRNet-W48[13]	76.3	90.8	82.9	81.2
HigherHRNet[30]	70.5	89.3	77.2	—
DEKR[31]	71.0	89.2	78.0	76.7
YOLO-Pose[32]	69.4	90.2	76.1	75.9
HR-Former-B[33]	75.6	90.8	82.5	80.8
MIPNet[34]	77.6	94.4	85.4	80.6
AD-HNN[35]	75.9	90.6	82.7	81.0
DPIT[36]	74.6	91.9	82.1	79.9
MogaNet-S[37]	74.9	90.7	82.8	80.1
PPE[38]	75.7	90.3	76.3	—
PCNet	**78.5**	**94.9**	**85.8**	**82.8**

MogaNet-S、PPE 等，PCNet 均有 2%～3%的提升。在 COCO 数据集上的实验验证了 PCNet 在常规场景，即非人群拥挤、遮挡严重的场景下仍然具备姿态估计有效性，可以达到比较理想的姿态估计效果。

在 CrowdPose 数据集上，采用和 COCO 数据集上的实验相同的输入图像尺寸和超参数配置。实验评价指标除上文介绍过的 AP、AP：50，AP：75 和 AR 之外，CrowdPose 数据集还根据其提出的密集度 CI 值对图像进行分类，CI 值小于 0.1 的图像属于 Easy 类图像；CI 值大于 0.1 且小于 0.8 的图像属于 Medium 类；CI 值大于 0.8 且小于 1 的图像为 Hard 类图像。AP^E、AP^M 和 AP^H 指标分别表示 Easy 类、Medium 和 Hard 类图像的平均准确率均值 mAP 指标。具体实验结果如表 2 所示。

表 2　CrowdPose 数据集实验结果对比

算法	AP	AP：50	AP^E	AP^M	AP^H
HRNet-W48[13]	71.3	91.1	80.5	71.4	62.5
HigherHRNet[30]	65.9	86.4	73.3	66.5	57.9
DEKR[31]	65.7	85.7	73.0	66.4	57.5
PINet[39]	68.9	88.7	75.4	69.6	61.5
TransPose-H[22]	71.8	91.5	79.5	72.9	62.2
HR-Former-B[33]	72.4	91.5	80.0	73.5	62.4
MIPNet[34]	72.8	92.0	80.6	73.1	65.2
AD-HNN[35]	70.8	86.5	77.0	68.5	59.6
I2R-Net[40]	72.3	92.4	79.9	73.2	62.8
PCNet	**75.4**	**93.1**	**83.2**	**76.5**	**67.1**

由表 2 可知，对比其他方法 TMI-Net 有较大的提升，AP^M 和 AP^H 指标均有大幅提高。对比基准网络 HRNet，本文提出的方法在相同的权重参数下平均精确度均值 AP 也有大约 4%的提升，其中 Hard 部分的实验数据结果提升最为明显，大约提升了 4.6%。HR-Former-B 是参考了 HRNet 的高分辨率网络结构提出的一种基于 Transformer 的姿态估计网络，相比之下本章的 TMI-Net 平均精度 AP 提高了 3%，准确说明了本论文提出的方法对于人群密集、遮挡严重的复杂场景有较好的实验效果，充分验证了 PCNet 在复杂场景下的人体姿态估计依然达到了理想效果。

由于 CrowdPose 数据集的实验数据更符合复杂场景的要求，因此本章通过基于 CrowdPose 数据集的消融实验来验证本文中对模型各模块改进的有效性。本章的消融实验将对模型中高分辨率网络的基于特征融合的 Transformer 编码器以及基于时空注意力机制的姿态矫正模块和 OHKM 损失函数对模型的效果影响展开实验和分析。本章的消融实验采用和上述对比实验相同的超参数配置，消融实验的结果如表 3 所示。

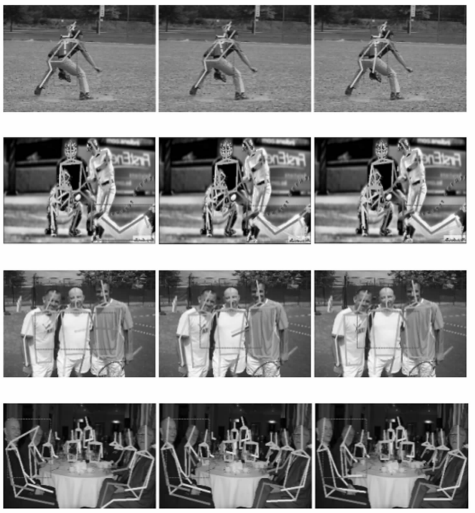
表 3　CrowdPose 数据集各模块消融实验

算法	AP	AP：50	AP^E	AP^M	AP^H
基准网络	71.3	91.1	80.5	71.4	62.5
＋OHKM	72.2	92.0	80.6	72.6	64.2
＋Encoder	72.3	91.9	80.9	72.6	64.2
＋姿态矫正	74.2	92.5	81.2	75.6	65.2
＋姿态矫正＋OHKM	74.9	92.6	82.0	76.1	66.7
PCNet	**75.4**	**93.1**	**83.2**	**76.5**	**67.1**

OHKM 表示基于在线困难点挖掘算法的损失函数，Encoder 表示基于特征融合的 Transformer 编码器，姿态矫正表示基于时空注意力机制的姿态矫正模块，通过实验结果可以看出，基于在线困难点挖掘算法的损失函数对模型效果大约有 0.9%的提升，相对于 PRF 和 HKRL 提升效果相对较小，基于特征融合的 Transformer 编码器的加入对模型效果产生了一定的优化，提升了大约 1.0%，姿态矫正模块作用比较明显，大约提升了 3.6%，同时包含 OHKM 损失函数的 HKRL 对 Hard 部分的数据精度也有较为明显的提升，精度大约增加了 4.2%。从实验结果可以看出，随着各个模块的引入，模型的效果呈现逐步提高的趋势；同时，姿态矫正模块的加入对模型的效果有较大提升。

2.4　可视化

图 8 展示了 PCNet 与其他方法的可视化结果对比，图片来自 CrowdPose 数据集，展示样例包含了人体遮挡和人群拥挤的复杂场景，子图虚线框突出了姿态差别，图 8（a）为 HRNet 可视化结果，图 8（b）为 MIPNet 可视化结果，图 8（c）为 PCNet 可视化结果。图 8 的第一行图片展示了人体自遮挡情况下，PCNet 成功识别到其他方法没识别的人体左臂关键点，图 8 的第二行图片展示了本章方法对比其他方法正确识别了腿部关键点，图 8 的第三行图片展示了本章方法对比其他方法成功识别出中间人体被遮挡的两个手臂，图 8 的第四行图片展示了本章方法对比其他方法识别到的遮挡关键点更多。



(a) HRNet (b) MIPNet (c) PCNet

图 8 姿态估计效果可视化对比

3 结束语

对于人群拥挤、肢体遮挡的复杂场景，本文在 HRNet 基础上，加入融合全局和局部信息的 Transformer 结构特征编码模块，引入到姿态估计模型的特征提取骨干网络中。通过该模块，模型可以更好地利用上下文信息，提高在人群拥挤、遮挡等复杂场景下的姿态估计精度；同时，在姿态矫正阶段，对于图像中的遮挡困难点，提出了一种基于时空注意力机制和级联结构的姿态矫正方法，对复杂场景中预测误差较大的关键点进行进一步修正，从而达到在自遮挡或互遮挡的情况下，关键点的估计依然可以保持较高的精度。在 COCO 以及 CrowdPose 数据集上取得了目前最优的估计效果。后续研究将关注在网络的推理速度和计算速度方面，可以通过对网络剪枝、压缩和量化的方式减少参数量，使得网络能够进一步满足实时性要求。

参考文献：

[1] DUBEY S, DIXIT M. A comprehensive survey on human pose estimation approaches [J]. Multimedia Systems, 2023, 29 (1): 167 - 195.

[2] 李佳宁，王东凯，张史梁. 基于深度学习的二维人体姿态估计：现状及展望 [J]. 计算机学报，2024，47 (1)：231 - 250.

[3] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks [J]. Communications of the ACM, 2017, 60 (6): 84 - 90.

[4] TOSHEV A, SZEGEDY C. Deeppose: human pose estimation via deep neural networks [C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014: 1653 - 1660.

[5] ZHENG C, WU W, CHEN C, et al. Deep learning-based human pose estimation: a survey [J]. ACM Computing Surveys, 2023, 56 (1): 1 - 37.

[6] 张 宇，温光照，米思娅，等. 基于深度学习的二维人体姿态估计综述 [J]. 软件学报，2021，33 (11)：4173 - 4191.

[7] 马双双，王 佳，曹少中，等. 基于深度学习的二维人体姿态估计算法综述. 计算机系统应用，2022，31 (10)：36 - 43.

[8] HE K, GKIOXARI G, DOLLAR P, et al. Mask r-CNN [C] //Proceedings of the IEEE international Conference On Computer Vision, 2017: 2961 - 2969.

[9] PAPANDREOU G, ZHU T, KANAZA N, et al. Towards accurate multi-person pose estimation in the wild [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, Hawaii: IEEE Press, 2017: 4903 - 4911.

[10] FANG H S, XIE S, TAI Y W, et al. Rmpe: regional multi-person pose estimation [C] //Proceedings of the IEEE International Conference on Computer Vision, 2017: 2334 - 2343.

[11] FANG Y, ZHAN B, CAI W, et al. Locality-constrained spatial transformer network for video crowd counting [C] //2019 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2019: 814 - 819.

[12] HUANG S, GONG M, TAO D. A coarse-fine network for keypoint localization [C] //Proceedings of the IEEE International Conference on Computer Vision, 2017: 3028 - 3037.

[13] SUN K, ZHAO Y, JIANG B, et al. High-resolution representations for labeling pixels and regions [J]. Arxiv Preprint Arxiv: 1904.04514, 2019.

[14] ZHANG J, CHEN Z, TAO D. Towards high performance human keypoint detection [J]. International Journal of Computer Vision, 2021, 129 (9): 2639 - 2662.

[15] PISHCHULIN L, INSAFUTDINOV E, TANG S, et al. DeepCut: joint subset partition and labeling for multi person pose estimation [C] // Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2016: 4929 - 4937.

[16] CAO Z, SIMON T, WEI S E, 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.

[17] 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, 2019: 11977 - 11986.

[18] LUO Z, WANG Z, HUANG Y, et al. Rethinking the

heatmap regression for bottom-up human pose estimation [C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 13264 - 13273.

[19] TANG W, YU P, WU Y. Deeply learned compositional models for human pose estimation [C] //Proceedings of the European Conference on Computer Vision (ECCV), 2018: 190 - 206.

[20] NIE X, FENG J, ZHANG J, et al. Single-stage multi-person pose machines [C] //Proceedings of the IEEE/CVF international conference on computer vision. 2019: 6951 - 6960.

[21] MCNALLY W, VATS K, WONG A, et al. Rethinking keypoint representations: modeling keypoints and poses as objects for multi-person human pose estimation [C] // European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 37 - 54.

[22] YANG S, QUAN Z, NIE M, et al. TransPose: keypoint localization via transformer [C] //Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 11802 - 11812.

[23] XU Y, ZHANG J, ZHANG Q, et al. Vitpose: Simple vision transformer baselines for human pose estimation [J]. Advances in Neural Information Processing Systems, 2022, 35: 38571 - 38584.

[24] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C] //Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017: 6000 - 6010.

[25] MARMOLIN H. Subjective MSE measures [J]. IEEE Transactions on Systems, Man, and Cybernetics, 1986, 16 (3): 486 - 489.

[26] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft coco: common objects in context [C] //Computer Vision-ECCV 2014: 13th European Conference, Zurich, Switzerland, Proceedings, Part V 13. Springer International Publishing, 2014: 740 - 755.

[27] 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.

[28] ANDRILUKA M, PISHCHULIN L, GEHLER P, et al. 2D human pose estimation: new benchmark and state of the art analysis [C] //Proceedings of the IEEE Conference on computer Vision and Pattern Recognition. 2014: 3686 - 3693.

[29] WU J, ZHENG H, ZHAO B, et al. Large-scale datasets for going deeper in image understanding [C] //2019 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2019: 1480 - 1485.

[30] 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.

[31] GENG Z, 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, 2021: 14676 - 14686.

[32] MAJI D, NAGORI S, MATHEW M, et al. YOLO-Pose: enhancing YOLO for multi person pose estimation using object keypoint similarity loss [J] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2022: 2636 - 2645.

[33] YUAN Y, FU R, HUANG L, et al. HRFormer: high-resolution transformer for dense prediction [J]. Arxiv preprint arxiv: 2110. 09408, 2021.

[34] KHIRODKAR R, CHARI V, AGRAWAL A, et al. Multi-Instance pose networks: rethinking top-down pose estimation [C] //Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 3122 - 3131.

[35] XU X, ZOU Q, LIN X. Adaptive hypergraph neural network for multi-person pose estimation [C] //Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36 (3): 2955 - 2963.

[36] ZHAO S, LIU K, HUANG Y, et al. Dpit: Dual-pipeline integrated transformer for human pose estimation [C] // CAAI International Conference on Artificial Intelligence. Cham: Springer Nature Switzerland, 2022: 559 - 576.

[37] LI S, WANG Z, LIU Z, et al. Efficient multi-order gated aggregation network [J]. Arxiv Preprint Arxiv: 2211. 03295, 2022.

[38] DAS A, DAS S, SISTU G, et al. Deep multi-task networks for occluded pedestrian pose estimation [J]. Arxiv Preprint Arxiv: 2206. 07510, 2022.

[39] WANG D, ZHANG S, HUA G. Robust pose estimation in crowded scenes with direct pose-level inference [J]. Advances in Neural Information Processing Systems, 2021, 34: 6278 - 6289.

[40] DING Y, DENG W, ZHENG Y, et al. I2R-net: intra-and inter-human relation network for multiperson pose estimation [C] // International Joint Conference on Artificial Intelligence, 2022: 855 - 862.