文章编号：2095-4980(2024)07-0781-11

基于MobileViT改进的红外人体姿态估计算法

张文扬[1a]，徐召飞[2]，刘　晴[2]，王科俊*[1b]，岳广辉[3]，王水根[2]，尚在飞[4]

(1 哈尔滨工程大学 a.研究生院，山东 烟台 264000；b.自动化学院，黑龙江 哈尔滨 150001；
2.烟台艾睿光电科技有限公司，山东 烟台 264000；3.深圳大学医学部 生物医学工程学院，广东 深圳 518060；
4.陆装驻烟台地区军事代表室，山东 烟台 264000)

摘　要：人体姿态估计主要依赖于视觉图像信息捕获关节点从而获得肢体和躯干的全局姿态信息。目前，基于可见光的深度学习方法具备较高的检测精确度，但隐私泄露的风险限制了其实际应用。同成本的红外探测器虽更能突出人体目标，但因成像分辨力较低，图像质量差，导致检测精确度下降。受视觉Transformer的启发，本文引入MobileViT-FPN提取人体关键点，利用MobileViT捕捉局部关节点特征和全局关节点特征关系，然后使用固定模式噪声(FPN)在多尺度上聚合这些表征信息，结合改进的OpenPose对关键点进行聚类，输出估计结果。在关键点级联阶段，注意力机制使模型自适应关注感兴趣区域，增强对遮挡部位的恢复。实验表明，该方法可以实时检测变化尺度和部分遮挡的红外人体目标，准确描绘人体姿态。

关键词：红外人体姿态估计；MobileViT主干网络；OpenPose网络；固定模式噪声

中图分类号：TN919.8　　**文献标志码：**A　　**doi：**10.11805/TKYDA2022149

Improved infrared human pose estimation algorithm based on MobileViT

ZHANG Wenyang[1a]，XU Zhaofei[2]，LIU Qing[2]，WANG Kejun*[1b]，YUE Guanghui[3]，
WANG Shuigen[2]，SHANG Zaifei[4]

(1a.Graduate School of Harbin Engineering University(Yantai)，Yantai Shandong 264000，China；1b.College of Automation，Harbin Engineering University，Harbin Helongjiang 150001，China；2.Yantai Arrow Photoelectric Technology Co.，Ltd.，Yantai Shandong 264000，China；3.School of Biomedical Engineering，Health Science Center，Shenzhen University，Shenzhen Guangdong 518060，China；4.Military Representative Office of Luzhuang in Yantai，Yantai Shandong 264000，China)

Abstract：Human pose estimation primarily relies on capturing joint points from visual image information to obtain global posture information of limbs and torso. Currently, depth learning methods based on visible light have high detection accuracy, but the risk of privacy leakage limits their practical application. Infrared detectors of the same cost can highlight human targets more effectively, but due to their lower imaging resolution and poor image quality, the detection accuracy is reduced. Inspired by visual Transformers, this paper introduces MobileViT-FPN to extract key human body points, using MobileViT to capture the relationship between local and global joint features, and then using Fixed Pattern Noise (FPN) to aggregate these representational information at multiple scales. Combined with an improved OpenPose for key point clustering, the estimated results are outputted. In the key point cascading phase, the attention mechanism allows the model to adaptively focus on the area of interest, enhancing the recovery of occluded parts. Experiments show that this method can real-time detect infrared human targets with varying scales and partial occlusions, accurately depicting human posture.

Keywords：infrared human pose estimation；MobileViT；OpenPose；Fixed Pattern Noise

人体姿态估计是以人体骨骼关节点为研究对象，通过检测关节点的位置信息，估计关节点之间的联系进而重构人体肢干的方法，在人机交互、智能安防、智能看护、虚拟现实和行为识别等应用中发挥关键作用[1-2]。随

收稿日期：2022-08-09；**修回日期：**2022-10-12
基金项目：国家自然科学基金资助项目(ZR202103080141)
*通信作者：王科俊　email:wangkejun@hrbeu.edu.cn

着摄像头等硬件监控设备和视觉智能分析平台的落地和推广，人们对人体姿态估计算法提出了更高的精确度和速度要求。

人体姿态估计主要分为单人姿态估计和多人姿态估计两种。单人姿态估计[3-6]主要通过基于坐标回归、基于热图检测等方法获取人体的关节点坐标；多人姿态估计由于要区别不同人体的关节点，需要在单人姿态估计的基础上添加额外的策略作为指导。鉴于多数实际场景常常需对多人的行为动作同时进行识别和判断，本文主要对多人姿态估计算法进行研究。多人姿态估计有两种思路：自顶向下和自底向上。自顶向下[7-10]的人体姿态估计方法分为两步：首先利用目标检测算法生成人体提议框；然后在每一个提议框内进行单人姿态估计。如 CHEN 等[7]提出级联金字塔网络，采用多阶段的策略构建单人人体姿态检测器。FANG 等[8]构建了专门用于处理人体提议框的空间变换网络，通过单独训练空间变换网络可以在不精准的框中提取到更高质量的提议框。与自顶向下的方法相反，自底向上[11-16]的人体姿态估计算法先利用单人姿态估计器检测图片中所有的人体关节点，再将不同人体的关节点聚成一类并拼接在一起。如，CAO 等[14]提出 OpenPose 网络，该网络使用部件亲和场学习人体肢体的向量场，对全局上下文进行编码，贪婪的自底向上解析，在实现实时性能的同时保持高准确性。CHENG 等[16]提出 Higher-HRNet，从一个高分辨力的子网络作为第一阶段开始，逐步添加从高分辨力到低分辨力的子网络，并将多个分辨力的子网络并行连接，从而产生丰富的高分辨力表征。

自顶向下的方法赖于人体提议框，单独估计每个人的姿势，通常需要大量计算，并不是真正的端到端系统。相比之下，自底向上的方法一次性检测图像中所有的关节点，只进行一次整体图像特征提取，即使人体数目增加也不会导致计算量翻倍，具备更高的效率和更小的模型规模。由于红外图像信噪比低，目标纹理特征少，分辨力低，上述基于可见光的姿态估计算法难以有效迁移至红外数据上。此外，以往自底向上的方法主要关注对关键点进行分组，并简单地使用单一分辨力的特征图，无法解决人体尺度变化大和人体遮挡的问题。本文调研发现，为应对人体尺度变化大的挑战，需要生成一个高分辨力且语义信息丰富的特征图；为解决遮挡问题，需要使网络突出关节点特征，抑制无用的特征信息。

针对于红外图像分辨力低、姿态尺度变化大、姿态自遮挡等问题，本文提出了基于 MobileViT[17]改进的红外人体姿态估计算法。

1 红外姿态数据集

本文创建了一个红外行人姿态数据集，数据集开源在艾睿光电的开源数据平台上。采用分辨力大小为 640×512 的手持红外摄像机采集客厅、天台、过道等多个场景下的红外图像数据，包含行走、跌倒、打电话等动作，如图 1 所示。该数据集包含 8 085 张红外图像，标记了 10 932 个人体目标。

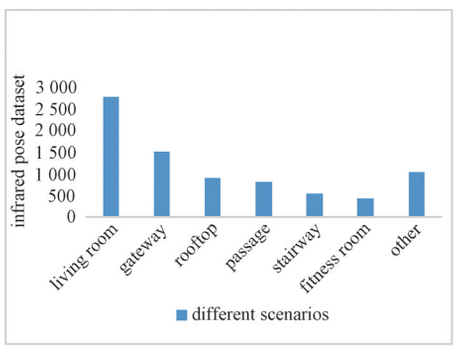


Fig.1 Distribution of datasets proposed in this paper
图 1 本文所提数据集的分布

姿态数据集主要对人体关节点和人体掩膜进行标注。关节点分为 3 个维度：关节点类别、关节点的二维坐标和关节点状态。其中关节点类别分为 17 类，用 0~16 的数字标号分别对应鼻子、左右眼、左右耳等人体主要关节点，具体对应关系如图 2 所示。关节点二维坐标以该关节在全图中的绝对位置进行表示；关节状态表示该关节的遮挡情况：0 代表该点并未出现在图中，1 代表该关节点在图中出现但被遮挡，2 代表该关节点出现了同时可标注无遮挡。人体掩膜标注了人体的外形轮廓，采用多边形格式标注，用一系列表示多边形各个顶点的二维坐标勾勒出目标轮廓。标注样例如图 3 所示，其中左图为原图，右图为标注图。在标注的过程中，对关节点小于 3 个、人体轮廓所占面积小于 100 个像素以及镜面中出现的人体均不标注其关节点，仅标注掩码以避免其参与损失函数的计算，如图 4 所示。其中左图为人体关节点小于 3 个，中图为人体像素小于 100 pixel，右图为镜面的人体。所有数据标注结果均仿照 COCO 数据集格式用 JSON 文件保存。

2 多人姿态估计算法

本文方法旨在以红外图像作为输入，定位图像中所有人的关节点和肢体位置。如图 5 所示，本文提出的网络结构由主干网络 MobileViT-FPN 和改进的 OpenPose(Improved OpenPose，I-OpenPose)组成。在 I-OpenpPose 网络中，每个阶段都输出中间监督细化的部件亲和场(Part Affinity Field，PAF)和关节点置信图(Part Confidence Maps，

PCM)。具体而言，首先通过主干网络提取红外图像的特征 $\boldsymbol{F}\in\mathbf{R}^{C\times W\times H}$($C$ 表示特征的通道数，W 表示特征的宽，H 表示特征的高)，作为 I–OpenPose 网络的输入；其次通过 I–OpenPose 网络先生成部件亲和场 L 与编码肢体之间的关联度：$L=(L_1,L_2,\cdots,L_c),c\in\{1\ 2\cdots B\}$ 表示所有肢体的 PAF，$L_b\in\mathbf{R}^{W\times H\times 2}$ 表示肢体 b 的 PAF，再生成关节点置信图 S 预测关节点的位置：$S=(S_1,S_2,\cdots,S_j),j\in\{1\ 2\cdots J\}$ 表示所有关节点的 PCM，$S_j\in\mathbf{R}^{W\times H}$ 表示关节点 j 的 PCM；最后，PAF 与 PCM 相结合，通过贪婪分析算法输出图像中所有人的关节点和肢体位置。

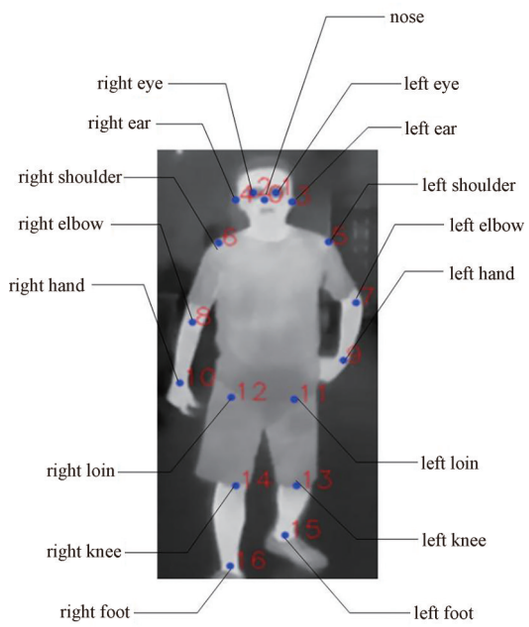


Fig.2 Diagram of human joints
图2 人体关节点示意图

Fig.3 Infrared pose dataset annotation sample
图3 红外姿态数据集标注样例

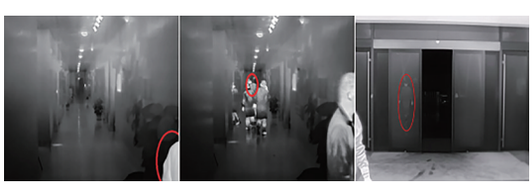


Fig.4 Annotation notes image
图4 标注注意事项示意图

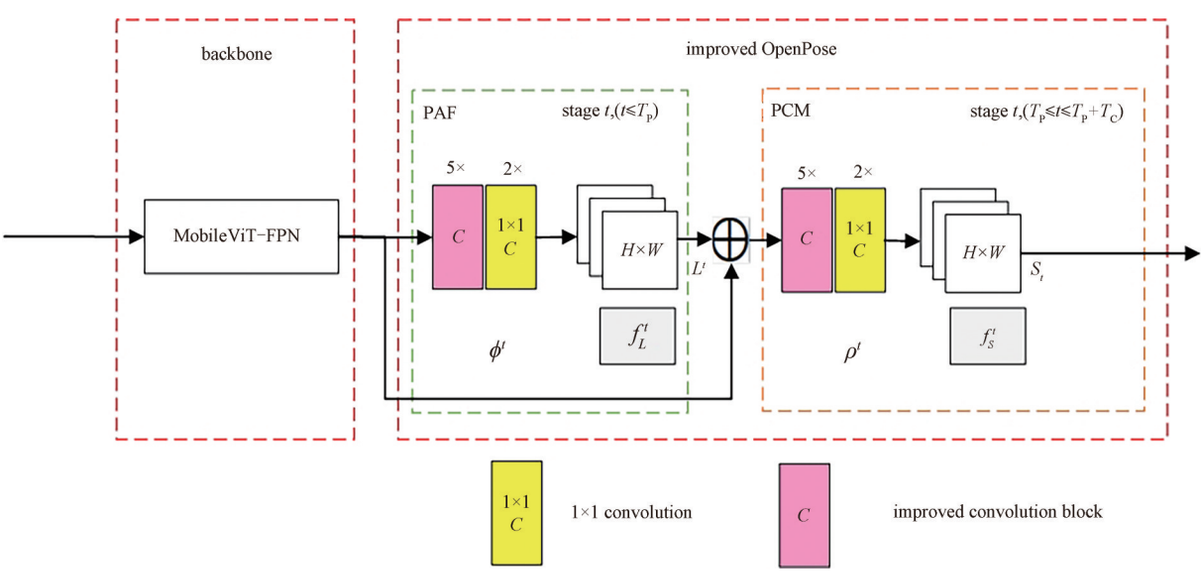


Fig.5 Structure of improved infrared human pose estimation algorithm based on MobileViT
图5 基于MobileViT改进的红外人体姿态估计算法结构图

2.1 主干网络

为解决识别过程中姿态尺度变化大的问题，设计了一个 MobileViT–FPN 主干网络，如图 6 所示。MobileViT 提取红外图像在不同尺度下的特征，FPN 聚合不同尺度的特征，得到一个分辨力高且语义信息丰富的特征，作为 I–OpenPose 网络的输入。

2.1.1 MobileViT

MobileViT 结合了卷积神经网络(空间归纳偏置和对数据增强的敏感性较低)和视觉 Transformer(输入自适应加

权和全局处理)的优点，能以较少的参数编码图像的局部和全局信息。MobileViT 由 MobileNetV2 块[18]和 MobileViT 块组成，如图 6 所示。其中，MobileNetV2 块旨在对特征图进行下采样，MobileViT 块旨在用较少的参数建模局部和全局信息。

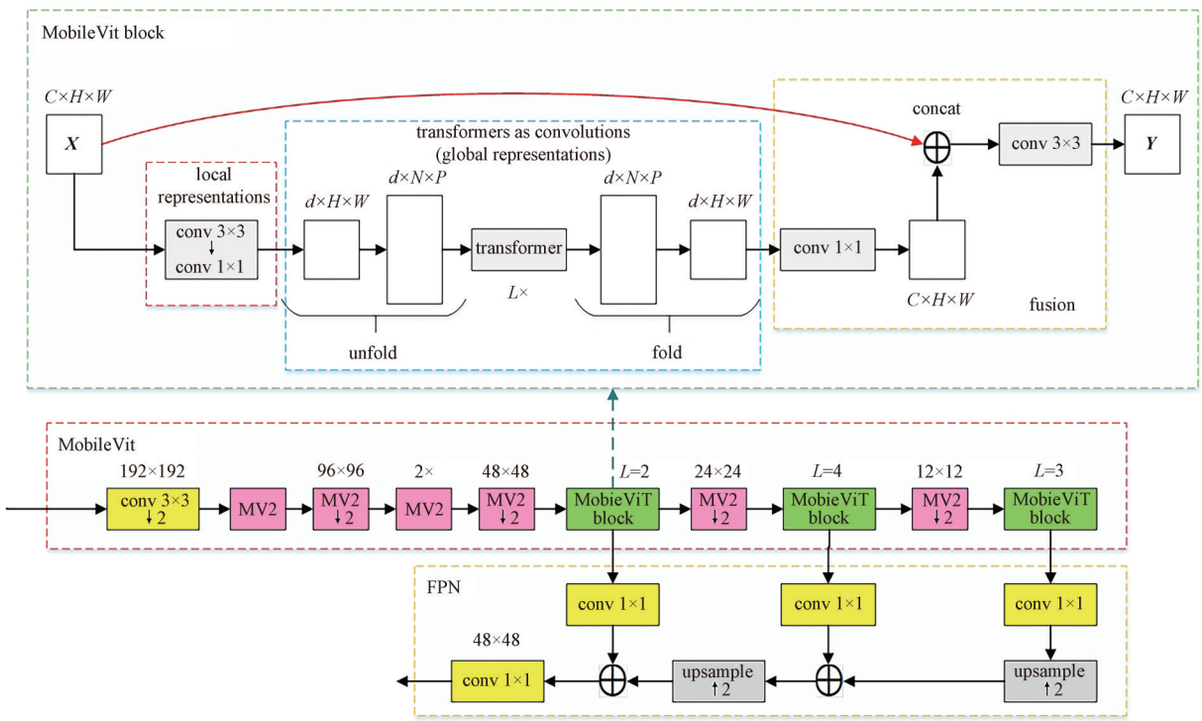


Fig.6 MobieViT–FPN network structure

图 6 MobieViT–FPN 网络结构图

对于给定的红外图像 $\boldsymbol{X}\in\mathbf{R}^{C\times H\times W}$，MobileViT 块首先通过 $n\times n$ 标准卷积编码局部信息，并使用 1×1 逐点卷积将输入通道 C 投影到高维空间 d，得到 $\boldsymbol{X}_L\in\mathbf{R}^{d\times H\times W}$；其次，为学习具有空间归纳偏置的全局特征，通过 unfold 操作将 $\boldsymbol{X}_L$ 变换到 $\boldsymbol{X}_U\in\mathbf{R}^{d\times P\times N}$，其中 $P=wh$，$N=\frac{HW}{P}$，N 表示将图片分成 N 个 patch 块，w 和 h 分别表示 patch 块的宽和高；使用 Transformer 编码每个 patch 之间的特征，得到 $\boldsymbol{X}_G\in\mathbf{R}^{d\times P\times N}$，通过 fold 操作将 $\boldsymbol{X}_G$ 变换到 $\boldsymbol{X}_F\in\mathbf{R}^{d\times H\times H}$；然后，通过 1×1 逐点卷积将输入通道 d 投影到低维空间 C，并将得到的结果与输入 $\boldsymbol{X}$ 拼接；最后，通过 $n\times n$ 标准卷积融合局部和全局特征。

$$\boldsymbol{X}_G(p)=\text{Transformer}(\boldsymbol{X}_U(p)),1\leqslant p\leqslant P \tag{1}$$

2.1.2 FPN

MobileViT 的最高层特征维度是输入维度的 1/32，高层的特征包含了丰富的语义信息，但很难准确保存关节点的位置信息；低层的特征语义信息较少，但可以准确包含关节点位置信息。为将低层的特征和高层的特征联系起来，得到一个识别和定位都准确的特征表示，本文采用 FPN[19]聚合 MobileViT 下采样倍数分别为 32 倍、16 倍、8 倍的特征。FPN 的网络结构如图 6 所示。

2.2 I–OpenPose

为实现更精确的预测，I–OpenPose 采用多阶段级联的方式细化 PAF 和 PCM，如图 5 所示。首先将主干网络的输出特征图作为第一阶段的输入，得到第一阶段的 PAF，在随后的每个阶段中，将前一阶段的预测和原始图像特征图连接起来，产生精确的预测。上述操作可表示为：

$$L^1=\phi^1(F) \tag{2}$$

$$L^t=\phi^t(F,L^{t-1}),\forall 2\leqslant t\leqslant T_{\text{P}} \tag{3}$$

式中：ϕ^t 为第 t 阶段推理的卷积神经网络(Convolutional Neural Networks，CNN)；T_{P} 为 PAF 总阶段数；L^t 为第 t 阶段的 PAF。

$$S^{T_P}=\rho^{t}(F,L^{T_P}),\forall t=T_P \tag{4}$$

$$S^{t}=\rho(F,L^{T_P},S^{t-1}),\forall T_P\leqslant t\leqslant T_P+T_C \tag{5}$$

式中：ρ^t 为第 t 阶段进行推理的 CNN；T_C 为 PCM 阶段的总数量；S^t 为第 t 阶段的 PCM。

2.2.1 卷积块

为抑制无用的特征信息，恢复遮挡的身体部位，并区分模糊的背景，在 OpenPose 每个阶段的卷积块引入 RSDANet，从而更好地模拟各个部位之间的空间关系，使模型自适应关注感兴趣的区域，如图 7 所示。

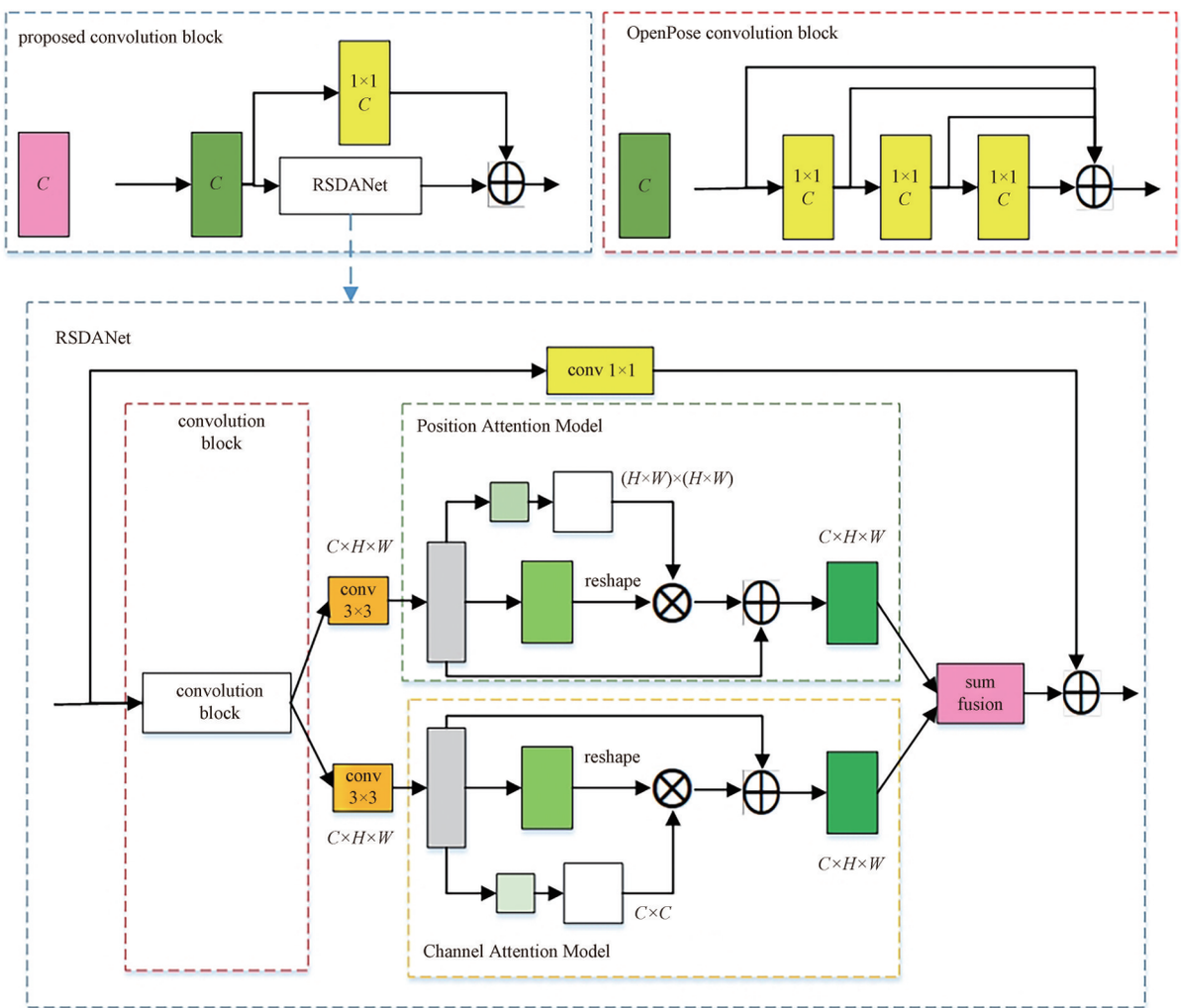


Fig.7 Convolution block network structure

图 7　卷积块网络结构图

RSDANet 模块包含位置注意力机制模块(Position Attention Model，PAM)、通道注意力模块(Channel Attention Model，CAM)、残差结构和光滑最大单元(Smooth Maximum Unit，SMU)激活函数[20]。本文在 OpenPose 卷积块上附加 PAM 和 CAM，分别对空间维度和通道维度上的语义依赖关系进行建模：首先，PAM 使用自注意力机制捕获特征图在任意 2 个位置之间的空间依赖关系，通过加权求和对所有位置的特征进行聚合更新；其次，CAM 使用自注意力机制捕获任意 2 个关节点之间的依赖关系，通过加权求和对所有关节点特征进行聚合更新[21]；最后，使用残差网络[22]融合 PAM 和 PCM 的特征，得到更丰富的特征表示。为提高网络性能，RSDANet 模块中所有的卷积操作后都采用 SMU 激活函数。

2.3 损失函数与推理过程

2.3.1 损失函数

本文采用中间监督训练的方式细化每一阶段的 PAF 和 PCM，得到更精确的输出，即在每一阶段结束使用 L2 损失函数，最后的损失为所有阶段损失之和。

$$f_L^t=\sum_{b=1}^{B}\sum_{p}W(p)\left\|L_b^t(p)-L_b^*(p)\right\|_2^2 \tag{6}$$

$$f_S^t=\sum_{j=1}^{J}\sum_{p}W(p)\left\|S_j^t(p)-S_j^*(p)\right\|_2^2 \tag{7}$$

$$f=\sum_{t=1}^{T_p}f_L^t+\sum_{t=T_p+1}^{T_p+T_C}f_S^t \tag{8}$$

式中：B为肢体数量；J为关节数量；L_b^*为真实PAF；S_j^*为真实PCM；L_b^t为预测输出PAF；S_j^t为预测输出PCM；$W(p)$为一个二元掩码，取0和1，$W(p)=0$表示像素p不参与损失计算；f_L^t为PAF的损失函数；f_S^t为PCM的损失函数；f为总体损失函数。

2.3.2 推理过程

推理过程主要分3步：a) 根据PCM定位关节点的位置；b) 使用PAF并联合PCM生成所有人的肢体；c) 使用贪婪分析算法[14]将肢体聚类到不同的人实例上。

PCM表示图片中关节点位置，共19个通道，对应18个关节点和背景。若图像中只有1个人，每个PCM只有1个峰值，如式(9)所示；若图像中有多人时，每1个人都对应1个可见关节点的峰值，如式(10)所示。

$$S_{j,k}^*(p)=\exp\left(-\frac{\left\|p-x_{j,k}\right\|_2^2}{\sigma^2}\right) \tag{9}$$

$$S_j^*(p)=\max_k S_{j,k}^*(p) \tag{10}$$

式中：$S_{j,k}^*$为第k个人第j个关节在像素p的真实PCM；$x_{j,k}$为第k个人第j个关节点的坐标；S_j^*为所有人第j个关节点在像素p的真实PCM；σ为控制峰值大小的数值。

PAF表示每个肢体的二维向量场，负责描述肢体的位置和方向信息，共38个通道，对应每个肢体向量场。如果图像只有1个人，每个PAF只有1个人的肢体向量，如式(11)~式(12)所示。式(13)~式(14)是判断图片上某点是否在肢体上；当图像中有多人时，则对应所有人在图片上某点处的肢体向量平均值，如式(15)所示。

$$L_{c,k}^*(p)=\begin{cases}v, & \text{if } p \text{ on the limb } c,k\\ 0, & \text{otherwise.}\end{cases} \tag{11}$$

$$v=(x_{j2,k}-x_{j1,k})/\left\|x_{j2,k}-x_{j1,k}\right\|_2 \tag{12}$$

$$0\leqslant v(p-x_{j1,k})\leqslant l_{c,k} \tag{13}$$

$$0\leqslant v_\perp(p-x_{j1,k})\leqslant \sigma_l \tag{14}$$

$$L_c^*(p)=\frac{1}{n_c(p)}\sum_k L_{c,k}^*(p) \tag{15}$$

式中：$L_{c,k}^*(p)$为第k个人第c个肢体在像素p的真实PAF值；v为归一化的肢体向量；$v_\perp$表示肢体的垂直向量；$l_{c,k}$为肢体的欧式距离；σ_l为肢体的宽度；$L_c^*(p)$为所有人在像素p的真实PAF值；$n_c(p)$表示所有人在像素p的非零向量的个数；$x_{j1,k}$表示第k个人第$j1$关节的坐标；$x_{j2,k}$表示第k个人第$j2$关节的坐标。

在推理阶段，通过式(16)计算肢体上2个关节点之间的PAF积分，判断肢体的关联情况。

$$E=\int_{u=0}^{u=1}L_c\big(p(u)\big)\frac{d_{j2}-d_{j1}}{\left\|d_{j2}-d_{j1}\right\|_2}\mathrm{d}u \tag{16}$$

$$p(u)=(1-u)d_{j1}+ud_{j2} \tag{17}$$

式中：$p(u)$为2个关节点之间的插值；d_{j1}和d_{j2}分别为肢体上2个关节点的坐标位置。

3　实验与结果

3.1 实验设置

使用本文创建的红外姿态数据集进行实验，所有数据按照9:1的比例划分为训练集和测试集。预处理阶段使用随机旋转([−40°,40°])、随机缩放([0.5,2.0])和随机裁剪的数据增强方案。网络输入图像分辨力为384×384，所有真实PCM、真实PAF和人体掩膜图分辨力都缩放至384×384。I–OpenPose网络中PAF的阶段数为4，PCM的阶段数为2，两者网络结构相同，但输入数据的通道数不同。第一阶段PAF的输入通道数为128，其余阶段通道数为166；第一阶段PCM的通道数为166，其余阶段通道数为185。使用AdamW优化器，权重衰减为0.01。学习率为4×10^{-4}，每迭代100轮，学习率下降10倍，学习率的最小值为2×10^{-5}，所有实验均训练300个epoch。

验证阶段将分辨力为640×512的红外图片放缩为分辨力为480×384的图片，不使用任何数据增强方案。在训练阶段，4个阶段生成的PAF和2个阶段生成的PCM都参与损失的计算；在验证阶段，只使用最后1个阶段生成的PAF和PCM完成姿态的推理。

3.2 评价准则

采用目标关节点相似度(Object Keypoint Similarity，OKS)评估预测值和真实值之间的误差，从而衡量姿态估计模型的性能。

$$S_{\mathrm{OK}}=\frac{\sum_i \exp\left(-\frac{d_i^2}{2S^2k_i^2}\right)\delta(v_i>0)}{\sum_i \delta(v_i>0)} \tag{18}$$

式中：d_i^2为预测值和真实值的欧几里得距离的平方；S^2为人体在图像中所占面积的平方；k_i^2为归一化因子，是数据集所有关节点标准差的平方；$v_i\in\{0,1,2\}$表示关键点的可见性，其中0表示不可见，1表示有遮挡，2表示无遮挡且可见。

平均精度(Average Precision，AP)表示所有OKS的平均精确度，$\mathrm{AP}^{0.50}$表示OKS阈值为0.5时的平均精确度，AP^{M}表示中等尺度目标的平均精确度，AP^{L}表示大尺度目标的平均精确度。

3.3 实验结果

表1为在红外姿态数据集上OpenPose与本文模型的实验结果。实验结果表明，本文所有的指标都优于自底向上的OpenPose，如AP提升了7.2个百分点。

表1 OpenPose与本文模型的指标对比(%)

Table1 Index comparison between OpenPose and proposed model(%)

model	AP	$\mathrm{AP}^{0.50}$	$\mathrm{AP}^{0.75}$	AP^{M}	AP^{L}
OpenPose	39.0	93.7	17.3	40.7	38.8
proposed	46.2	96.2	33.9	48.1	45.9

如图8所示，本文方法在实际场景中检测效果优于OpenPose，能够检出OpenPose无法检测和识别错位的关节点。此外，遮挡关节点的检测一直都是姿态识别的难点。从图8中可看出，本文所提模型能更好地处理身体部位自遮挡情况。

如图9所示，本文的模型有很好的场景泛化效果，不仅在不同的场景中能准确地识别出人体姿态，对于较小的人体目标也能有效识别。

3.4 消融实验

本节将通过消融实验验证主干网络和I–OpenPose网络的有效性。所有实验结果均来自所构造的红外姿态数据集。

3.4.1 主干网络

I–OpenPose网络保持不变，主干网络分别采用VGG19[23]、MobileNetV2、Lite–HRNet[24]和MobileVit，MobileViT–FPN在红外姿态数据集进行训练和测试，并采用OKS作为评价准则，评估结果如表2所示。

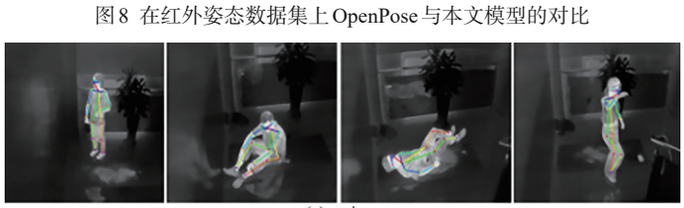
从表2中可以看出，本文的方法性能最好，VGG19性能次之，MobileNetV2和Lite–HRNet最差。VGG19是OpenPose的主干网络，本文的实验结果均与其进行对比，其AP为39.0%。MobileNetV2是一个轻量化的卷积神经网络，通过深度可分离卷积减少网络参数，提升速度，但降低了其特征提取能力，其AP仅为40.3%。Lite–HRNet通过高分辨力特征金字塔学习多尺度的特征，从而实现更精确的人体姿势估计，AP提升了2.2个百分点。人体两

两关节之间是相互联系的，如人跌倒时，需要多个关节点的配合。以上基于CNN的主干网络只能捕捉近距离关节点之间的联系，无法捕捉远距离关节之间的依赖关系，因此引入MobileViT聚合关节点之间局部和全局的特征后，其AP提升了3.4个百分点。最后采用FPN聚合MobileViT的不同尺度的特征，实验表明，其AP达到42.9%。

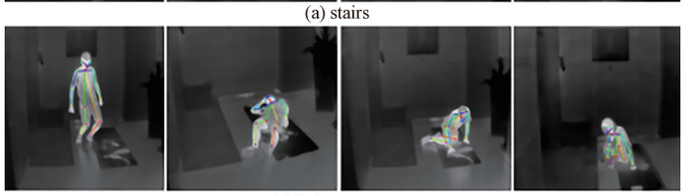
(a) OpenPose

(b) this model

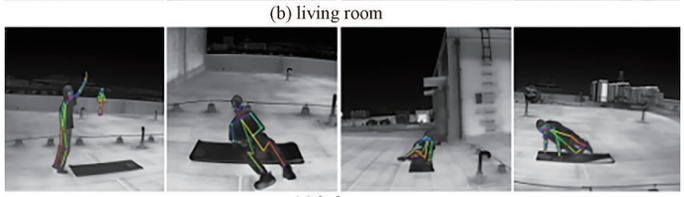
Fig.8 Comparison between OpenPose and the proposed model on the infrared pose dataset
图8 在红外姿态数据集上OpenPose与本文模型的对比



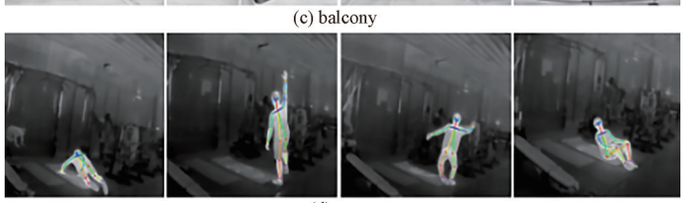
(a) stairs



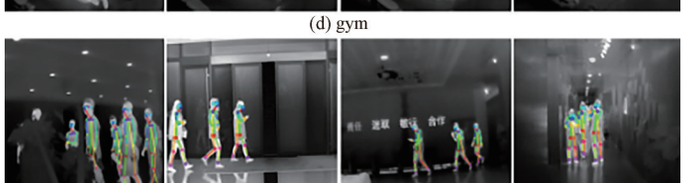
(b) living room



(c) balcony



(d) gym



(e) other complex scenes of multi-person detection

Fig.9 Detection effect of the proposed model in different scenes
图9 本文模型在不同场景的检测效果

3.4.2 多尺度特征

在主干网络和I–OpenPose网络不变情况下，分别实验了下采样倍数为16的MobileViT、下采样倍数为8的MobileViT和MobileViT–FPN(融合MobileViT下采样倍数为32、16和8的网络)对模型的影响，实验结果如表3所示。

表2 不同主干网络对模型的影响(%)

Table2 Effects of different backbone networks on model(%)

backbone	AP	$AP^{0.50}$	$AP^{0.75}$	AP^{M}	AP^{L}
VGG19	39.0	93.7	17.3	40.7	38.8
MobileNetV2	40.3	93.0	18.3	41.7	39.6
Lite–HRNet	41.2	94.9	19.8	43.2	39.7
MobileVit	42.4	93.7	24.1	43.4	42.5
MobileViT–FPN	42.9	94.9	24.1	44.9	42.5

表3 不同下采样倍数对模型的影响(%)

Table3 Effects of different down–sampling ratios on the model(%)

backbone	downsample	AP	AP^{M}	AP^{L}
MobileViT	8(48×48)	42.4	43.4	42.5
	16(24×24)	22.9	21.0	27.1
MobileViT–FPN	8(48×48)	42.9	44.9	42.5

下采样倍数大，得到特征图的分辨力小，虽然其语义信息丰富，但淹没了关节点的位置信息，下采样16倍的特征图AP仅有22.9%；下采样倍数小，得到特征图的分辨力大，虽考虑了关节点的位置信息但忽视了语义信息，下采样8倍的特征图AP达到42.4%。

3.4.3 注意力机制

为在I–OpenPose网络中使模型更关注关节点区域，抑制无用的背景，本文在每个阶段的卷积块后引入注意力机制，表4为不同注意力机制的实验结果。

表4 不同注意力机制对模型的影响(%)

Table4 Effects of different attention mechanisms on the model(%)

backone	attention	AP	AP^{M}	AP^{L}
MobileViT–FPN	—	42.9	44.9	42.5
	SENet	43.2	45.0	43.0
	SANet	44.0	45.9	43.5
	DANet	45.0	46.5	44.5
	RSDANet	46.2	48.1	45.9

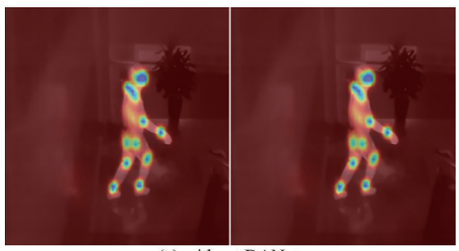
SENet[25]采用通道注意力机制将重要通道的特征强化，非重要通道的特征弱化，其AP比没引入之前提高了0.3个百分点。SANet[26]在空间和通道注意力机制的基础上，引入了特征分组与通道置换，其效果优于SENet。DANet[21]不仅通过通道注意力机制自适应加权通道特征，还通过空间注意力机制捕捉像素级别的依赖关系，其AP比没引入之前提高了2.1个百分点。从图10可知，引入DANet后，对于关节点的定位更加准确，甚至可以定位遮挡的关节点，如图10的左肘和左手所示。从图11可知，引入DANet之后，本文的模型在每个阶段都可以细化肢体的位置。

本文在DANet的基础上，引入残差结果，增加特征的多样性和加快训练，从表4可知，AP达到46.2%。实验结果表明，在PAM和PCM每个阶段的卷积块后引入RSDANet模块，不仅提高了模型精确度，还细化了PAF和PCM的输出，使模型更加关注关节点区域，改善对遮挡关节点的检测。

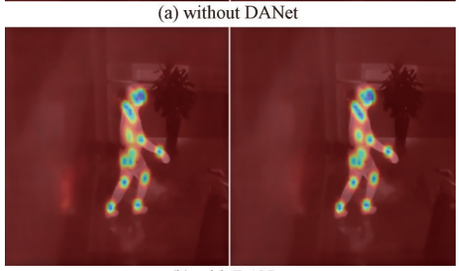
4 结论

本文提出了一种基于MobileViT的改进的红外人体姿态估计算法。首先，通过MobileViT–FPN得到分辨高且语义信息丰富的特征表示，不仅解决了人体姿态估计中尺度变化大的问题，还能提高对小目标姿态的估计；其次，在I–OpenPose网络中，引入RSDANet注意力机制，使网络自适应关注关节点区域，提高了对遮挡关节点的识别能力。

该算法仍存在不足：在一些特殊场景中检测效果差，比如人体姿势非直立向上；在拥挤场景中，容易出现关节点漏检的情况。在未来的研究工作中，需要着重解决在特殊场景中人体姿态的检测。

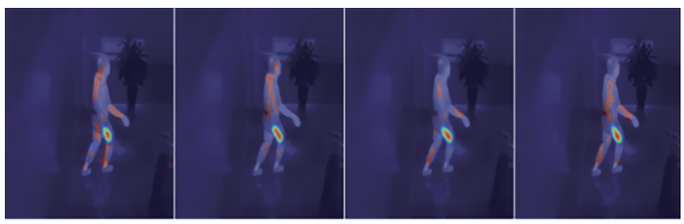


(a) without DANet

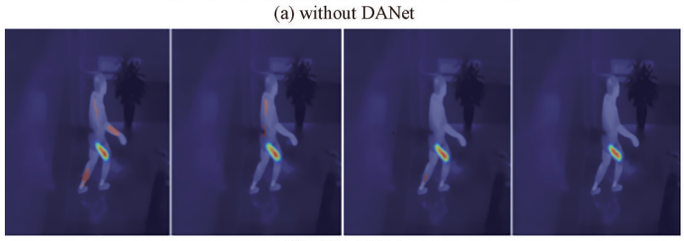


(b) with DANet

Fig.10 Heatmap of joints at different stages
图 10 不同阶段关节点的热力图



(a) without DANet



(b) with DANet

Fig.11 Heatmap of right thigh at different stages
图 11 不同阶段右大腿的热力图

参考文献：

[1] 乔迤,曲毅. 基于卷积神经网络的 2D 人体姿态估计综述[J]. 电子技术应用, 2021,47(6):15-21. (QIAO Yi,QU Yi. Overview of 2D human pose estimation based on convolutional neural network[J]. Application of Electronic Technique, 2021,47(6):15-21.) doi: 10.16157/j.issn.0258-7998.201087.

[2] 邓益侬,罗健欣,金凤林. 基于深度学习的人体姿态估计方法综述[J]. 计算机工程与应用, 2019(19):22-42. (DENG Yinong, LUO Jianxin,JIN Fenglin. Overview of human pose estimation methods based on deep learning[J]. Computer Engineering and Applications, 2019(19):22-42.) doi:10.3778/j.issn.1002-8331.1906-0113.

[3] TOSHEV A,SZEGEDY C. DeepPose:human pose estimation via deep neural networks[C]// 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus,OH,USA:IEEE, 2014:1653-1660. doi:10.1109/CVPR.2014.214.

[4] CARREIRA J,AGRAWAL P,FRAGKIADAKI K,et al. Human pose estimation with iterative error feedback[DB/OL]. (2015-07-23)[2022-08-09]. https://arxiv.org/abs/1507.06550. doi: 10.48550/arXiv.1507.06550.

[5] WEI S E,RAMAKRISHNA V,KANADE T,et al. Convolutional pose machines[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Las Vegas,NV,USA:IEEE, 2016:4724-4732. doi:10.1109/CVPR.2016.511.

[6] NEWELL A, YANG Kaiyu, DENG Jia. Stacked hourglass networks for human pose estimation[J/OL]. Springer International Publishing, 2016:483-499. doi:10.1007/978-3-319-46484-8_29.

[7] CHEN Yilun, WANG Zhicheng, PENG Yuxiang, et al. Cascaded pyramid network for multi-person pose estimation[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018: 7103-7112. doi: 10.1109/CVPR.2018.00742.

[8] FANG Haoshu, XIE Shuqin, TAI Y W, et al. RMPE: Regional Multi-person Pose Estimation[C]// 2017 IEEE International Conference on Computer Vision(ICCV). Venice,Italy:IEEE, 2017:2353-2362. doi:10.1109/ICCV.2017.256.

[9] SUN Ke, XIAO Bin, LIU Dong, et al. Deep high-resolution representation learning for human pose estimation[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Long Beach, CA, USA: IEEE, 2019: 5686-5696. doi: 10.1109/CVPR.2019.00584.

[10] LI W, WANG Z, YIN B, et al. Rethinking on multi-stage networks for human pose estimation[DB/OL]. (2019-01-01)[2022-08-09]. https://arxiv.org/abs/1901.00148. doi: 10.48550/arXiv.1901.00148.

[11] INSAFUTDINOV E, PISHCHULIN L, ANDRES B, et al. DeeperCut: a deeper, stronger, and faster multi-person pose estimation model[C]// The 14th European Conference. Amsterdam, the Netherlands: Springer International Publishing, 2016: 34-50. doi: 10.1007/ 978-3-319-46466-4_3.

[12] PISHCHULIN L, INSAFUTDINOV E, TANG Siyu, et al. DeepCut: joint subset partition and labeling for multi person pose estimation[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Las Vegas, NV, USA: IEEE, 2016: 4929-4937. doi:10.1109/CVPR.2016.533.

[13] NEWELL A, HUANG Zhi'ao, DENG Jia. Associative embedding: end-to-end learning for joint detection and grouping[C]//

Proceedings of the 31st International Conference on Neural Information Processing Systems. [S.l.]: Association for Computing Machinery, 2017:2274–2284.

[14] CAO Zhe, SIMON T, WEI S, et al. Realtime multi–person 2D pose estimation using part affinity fields[DB/OL]. (2016–11–24) [2022–08–09]. https://arxiv.org/abs/1611.08050. doi:10.48550/arXiv.1611.08050.

[15] KREISS S, BERTONI L, ALAHI A. PifPaf: composite fields for human pose estimation[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Long Beach, CA, USA: IEEE, 2019: 11969–11978. doi: 10.1109/CVPR.2019.01225.

[16] CHENG Bowen, XIAO Bin, WANG Jingdong, et al. HigherHRNet: scale–aware representation learning for bottom–up human pose estimation[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Seattle, WA, USA: IEEE, 2020: 5385–5394. doi:10.1109/CVPR42600.2020.00543.

[17] MEHTA S, RASTEGARI M. Mobilevit: light–weight, general–purpose, and mobile–friendly vision transformer[C]// IEEE Conference on Computer Vision and Pattern Recognition(CVPR). 2021. doi: 10.48550/arXiv.2110.02178.

[18] SANDLER M, HOWARD A, ZHU Menglong, et al. MobileNetV2: inverted residuals and linear bottlenecks[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Salt Lake City, UT, USA: IEEE, 2018:4510–4520. doi:10.1109/CVPR.2018.00474.

[19] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Honolulu, HI, USA: IEEE, 2017:936–944. doi:10.1109/CVPR.2017.106.

[20] BISWAS K, KUMAR S, BANERJEE S, et al. Smooth maximum unit: smooth activation function for deep networks using smoothing maximum technique[C]// IEEE Conference on Computer Vision and Pattern Recognition(CVPR). New Orleans, LA, USA: IEEE, 2021:784–793. doi:10.1109/CVPR52688.2022.00087.

[21] FU Jun, LIU Jing, TIAN Haijie, et al. Dual attention network for scene segmentation[C]// IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Long Beach, CA, USA: IEEE, 2019:3146–3154.

[22] HE Kaiming, ZHAGN Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]// IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Las Vegas, NV, USA: IEEE, 2016:770–778.

[23] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for Large–Scale image recognition[DB/OL]. (2014–09–04) [2022–08–09]. https://arxiv.org/abs/1409.1556. doi: 10.48550/arXiv.1409.1556.

[24] YU Changqian, XIAO Bin, GAO Changxin, et al. Lite–HRNet: a lightweight high–resolution network[C]// 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Nashville, TN, USA: IEEE, 2021: 10435–10445. doi: 10.1109/CVPR46437.2021.01030.

[25] HU Jie, SHEN Li, SUN Gang. Squeeze–and–Excitation networks[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018:7132–7141. doi:10.1109/CVPR.2018.00745.

[26] ZHANG Qinglong, YANG Yubin. SA–Net: shuffle attention for deep convolutional neural networks[C]// ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). Toronto, ON, Canada: IEEE, 2021: 2235–2239. doi:10.1109/ICASSP39728.2021.9414568.

作者简介：

张文扬(1998–)，男，在读硕士研究生，主要研究方向为姿态估计、行为识别.email:zwy1005713095@163.com.

徐召飞(1993–)，男，博士，工程师，主要研究方向为红外图像处理技术.

刘　晴(1995–)，女，硕士，工程师，主要研究方向为红外视觉的AI图像算法.

王科俊(1962–)，男，博士，教授，博士生导师，主要研究方向为模糊混沌神经网络、自适应逆控制理论、可拓控制、网络智能控制、多模态生物特征识别、联脱机指纹考试身份鉴别系统、微小型机器人系统.

岳广辉(1990–)，男，博士，副研究员，助理教授，主要研究方向为图像处理、计算机视觉、医学图像分析、机器学习及其应用的研究工作.

王水根(1990–)，男，博士，主要研究方向为人工智能、图像处理和机器学习.

尚在飞(1989–)，男，硕士，工程师，主要研究方向为图像处理.