DOI:10.19651/j.cnki.emt.2417194

基于增强特征融合的轻量级人体姿态估计网络*

施昕昕 张昊亮
(南京工程学院自动化学院 南京 211167)

摘 要:为了提高轻量化人体姿态估计网络对不同阶段特征图的信息提取和特征融合能力和关键点热力图与分类特征图的后处理能力,提出了一种基于多阶段多层级特征融合的人体姿态估计网络。首先设计了多层级特征融合模块,以提高神经网络模型对特征图的信息提取和归纳总结能力;接着设计了结合特征融合模块设计了特征融合分支,以达到保留模型不同阶段的信息不会随长期卷积运算而丢失的效果;最后对模型输出的关键点分类图进行后处理操作,对分类部分使用分类损失增强模块进行进一步增强,使其能够更好地专注于关键点分类任务,以提高模型输出的准确性。在 CrowdPose 数据集进行性能测试,本文算法和 LitePose 算法在 XS 结构下的 AP 值分别为 50.7%和 48.4%;在 S 结构下,AP 值分别为 59.1%和 58.3%。在 MS COCO val2017 数据集进行性能测试,本文算法和 LitePose 算法在 XS 结构下的 AP 值分别为 41.9%和 40.6%;在 S 结构下,AP 值分别为 57.0%和 56.8%。实验结果表明,本文算法提出的多层级特征融合模块和高分辨率融合分支以及后处理操作对人体姿态估计网络检测性能提升具有正向作用。

关键词:人体姿态估计;轻量级网络;多尺度特征融合;深度可分离卷积
中图分类号:TN911.73 **文献标识码**:A **国家标准学科分类代码**:520.6040

Light-weight human pose estimation network based on enhanced feature fusion method

Shi Xinxin Zhang Haoliang
(School of Automation, Nanjing Institute of Technology, Nanjing 211167, China)

Abstract: To enhance the lightweight human pose estimation network's ability to extract information and fuse features from different stages of feature maps, as well as improve the post-processing capability of keypoint heatmaps and classification feature maps, a human pose estimation network based on multi-stage and multi-level feature fusion is proposed. First, a multi-level feature fusion module is designed to improve the neural network model's ability to extract and summarize information from feature maps. Next, a feature fusion branch is designed in conjunction with the feature fusion module to ensure that information from different stages of the model is preserved without being lost due to long convolution operations. Finally, post-processing operations are applied to the model's output keypoint classification maps, utilizing a classification loss enhancement module for further enhancement, allowing the model to better focus on the keypoint classification task and improve the accuracy of its outputs. Performance testing is conducted on the CrowdPose dataset, where the AP values of the proposed algorithm and the LitePose algorithm under the XS structure are 50.7% and 48.4%, respectively; under the S structure, the AP values are 59.1% and 58.3%. Performance testing is conducted on the MS COCO val2017 dataset, where the AP values of the proposed algorithm and the LitePose algorithm under the XS structure are 41.9% and 40.6%, respectively; under the S structure, the AP values are 57.0% and 56.8%. Experimental results indicate that the multi-scale feature fusion module, high-resolution fusion branch, and post-processing operations proposed in this paper positively contribute to improving the detection performance of the human pose estimation network.

Keywords: human pose estimation; lightweight network; multiscale feature fusion; depth separable convolution

0 引 言

人体姿态估计是深度学习与计算机视觉领域当中一个具有重要价值的研究方向,其主要目标是从图像中辨析与定位人体主要关键节点所在的位置。人体姿态估计识别方法在进行行为识别[1-2]、人机交互[3]、安全防范[4-5]等领域有

收稿日期:2024-10-24
* 基金项目:国家自然科学基金面上项目(61873120)、南京工程学院校级科研基金创新基金重大项目(CKJA201903)资助

着广泛的应用前景。

基于深度学习的二维图像人体姿态估计方法可分为单人人体姿态估计和多人人体姿态估计两种。其中，Thshev 等[6]提出的单人人体姿态估计方法 DeepPose 是一种基于关键点的坐标回归的方法，然而其模型的泛化能力欠佳。Tompson 等[7]首次提出将热力图回归方式加入到人体姿态估计领域当中进行计算以提高预测结果精度。Newell 等[8]提出了 Hourglass 网络达到了人体姿态估计领域中预测准确度新的高度，在此基础上，Sun 等[9]以 Hourglass 网络为基础，增加了高分辨率分支图并提出了 HRNet 网络结构，该网络结构因为其优越的性能和准确率，成为了后续多种人体姿态估计网络设计基础的骨干网络。在多人人体姿态估计网络方面，旷视科技等[10]提出的级联金字塔网络使得姿态估计模型对于存在多种遮挡的情况的人体关键点预测结果有了进一步地提升。Cheng 等[11]提出的 HigherHRNet 将 HRNet 的网络结构进行了调整并转化运用到自下而上的多人人体姿态估计任务上。上述网络虽然在进行人体姿态估计上取得了较好的结果，但其庞大的参数量与计算量在边缘计算系统上难以部署和运用。

为了降低 HRNet 等网络的复杂度，使得其能够在算力较低的边缘计算系统部署，Wang 等[12]提出的 Small-HRNet 通过降低网络的宽度和深度对网络结构进行简化，但是这在降低复杂度的同时也导致了关键点检测的性能下降。Yu 等[13]提出的 Lite-HRNet 网络利用了高效的卷积神经网络替代类 Small-HRNet 当中的残差模块获得了更佳性能，谢唯嘉等[14]在 Lite-HRNet 上添加了空洞卷积与注意力模块进一步提高了准确度。但是其依然需要通过其他网络提取出单个人体实例之后才能进行姿态估计，与自下而上的识别模型相比耗时较长。Wang 等[15]提出的 LitePose 以 HigherHRNet 作为基准网络，削减了高分辨率分支并且用深度卷积替换了常规卷积层，成功实现了模型的轻量化，然而缺乏对高分辨率图的处理、对后续的卷积上采样、关键点分类图的后处理操作使得其识别准确率不高。

为了解决轻量级卷积神经网络 LitePose 对不同阶段不同尺度的特征图进行特征融合与低级细节特征图与高级语义特征图利用不充分的问题和模型输出的热力图与分类图在后处理上不充分的问题，本文提出了一种轻量级人体姿态估计网络 MFF-Pose，该网络以 LitePose 作为基准网络，设计并添加了多层级特征融合模块(multi-stage feature fusion module，MFF)以加强对不同分辨率、不同输出阶段的特征图融合的能力，同时设计了高分辨率特征融合分支结构(high resolution feature fusion，HRFF)以填补后阶段模型卷积上采样时对不同尺度、不同输出阶段的信息丢失，从而增强网络模型对不同阶段特征图的信息融合能力与信息挖掘能力，为反卷积上采样操作提供丰富的信息支持。此外，对 LitePose 的分类图输出，研究并设计了增强分类概括的模块(enhanced associated embedding，EAE)，对关键点分类图进行了全局的信息补充以及额外增强融合处理，提高了模型在关键点分类任务上的准确性。

1　相关工作

1.1　高分辨率网络

高分辨率网络在姿态估计领域取得了极佳的效果，HRNet 采用并行子网方式，通过将多阶段的多尺度特征融合以维持高分辨率特征图，有效地利用了不同尺度之间的特征信息，弥补了下采样导致空间分辨率损失对任务的影响。HigherHRNet 以 HRNet 为基础，将 HRNet 的输入输出进行了调整，使得高分辨率网络可以适用于自底向上的人体姿态估计任务。然而，模型始终维持高分辨率特征图，也加大了模型的整体复杂度，同时对高分辨率图进行处理带来了显著的计算量上升与运行内存占用问题，难以在边缘计算设备上进行部署。

对此，LitePose 将 HigherHRNet 的骨干网络当中的普通卷积替换成为了深度卷积，其设计的 LitePose 网络结构移除了复杂的高分辨率分支，在实现了模型的轻量化的同时维持了一定的识别准确率。然而其虽然解决了高分辨图带来的计算量和内存占用问题，却同时失去了高分辨率分支保留的细节信息，导致了网络整体精度下降。

1.2　特征融合

Huang 等[16]提出了一种稠密连接结构的网络模型 DenseNet，其结构如图 1(a)所示。该模型受到 He 等[17]提出的 ResNet 残差连接启发，将输入特征图的不同阶段的特征图进行拼接，一同送入下一个卷积层，这种稠密连接的方式有有助于增加梯度，让模型学习更多的信息，但这种结构显著增加了参数量与计算量。Lee 等[18]提出的 VoVNet 受到 DenseNet 稠密结构的启发并认为 DenseNet 当中对原始输入使用过于重复，存在信息冗余，并以此提出了将不同阶段的特征图只在最后拼接的如图 1(b)所示的模型结构，该结构在减少了参数数量和计算量的同时取得了比 DenseNet 更好的结果。Wang 等[19]提出的 CSPNet 结合了 DenseNet 当中的稠密链接方式，并认为对网络进行分组处理可以在维持精度的同时进行网络的参数和计算量的大幅度削减。Wang 等[20]提出的 ELAN 结构结合了 VoVNet 和 CSPNet 的结构样式，设计出了结合了多种梯度和分组的多尺度特征融合模块，在 Wang 等[21-22]提出的 YoloV7 和 YoloV9 的网络模型中取得了较好的效果。然而缺少轻量化卷积与注意力模块的使用使得该特征融合结构在轻量化网络方面存在着较多改进空间。

1.3　注意力机制

Hu 等[23]提出的 SENet 网络中的通道注意力机制(channel attention，CA)是极具代表性的一个模块，将通道注意力引入卷积块，学习每个卷积块的通道注意力，给深度卷积神经网络结构带来明显的提升。在给定输入特征后，CA 模块先对每个通道单独进行全局平均池化，再接着进

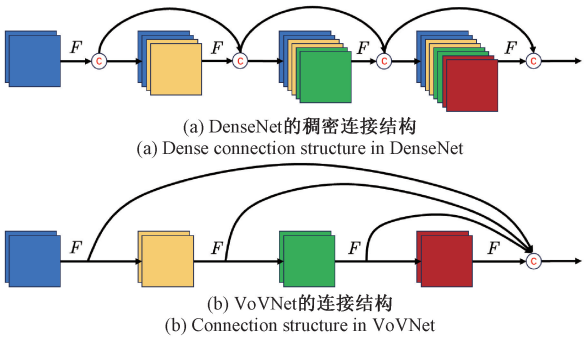


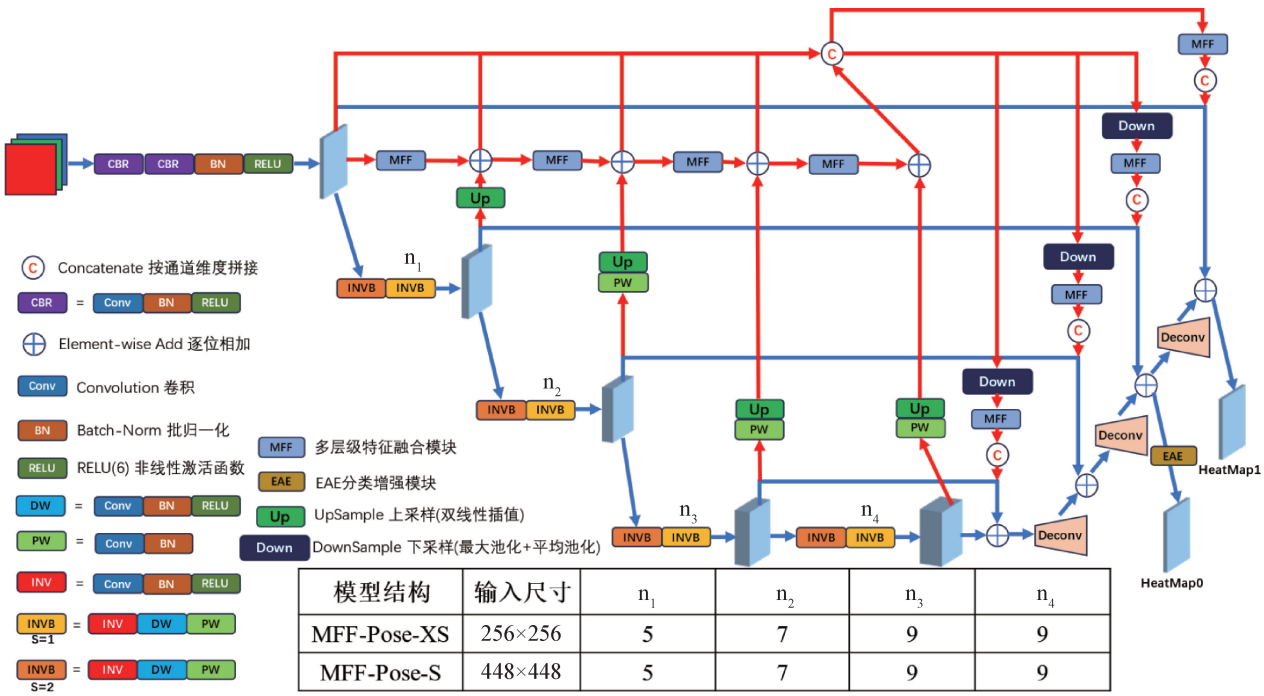
图 1 特征融合示意图

Fig. 1 Feature fusion blocks of DenseNet and VoVNet

行 2 个具有非线性的完整连接层,最后使用 Sigmoid 函数来生成通道权值。Woo 等[24]结合 CA 模块设计了卷积神经网络注意力机制组合模块(convolutional block attention module, CBAM),其中通道注意力模块(channel attention module, CAM)与 CA 模块类似,空间注意力模块(spatial attention module, SAM)通过将输入的特征图进行全局平均池化和最大池化,再通过卷积生成一个 $1*H*W$ 的注意力权重图,通过将权重和特征图进行相乘,提高了模型对局部区域的注意力。Wang 等[25]针对 CA 模块的对特征图维度压缩导致信息丢失的问题进行改进并提出了高效通道注意力模块(efficient channel attention, ECA),实验中得到了与 CA 模块相比更好的结果,同时更具轻量化。Hou 等[26]提出了坐标注意力机制(coordinate attention, CA),该注意力机制可以将特征图沿着 x 轴和 y 轴进行信息聚合和学习,生成像素粒度的权重图,将权重和特征图进行乘积可以提高模型对细节的学习和把握能力。将注意力机制融入特征融合模块,可以起到更好的效果。

2 增强特征融合的人体姿态估计网络设计

本文提出的人体姿态估计模型网络结构如图 2 所示,该网络以文献[27]中提出的 MobileNetV2 作为特征提取的骨干网络,按照特征图分辨率的不同分为 4 个层级。在最后的层级处理结束后,通过不断的反卷积上采样操作恢复特征图的分辨率以生成最终的预测热力图和关键点分类图。



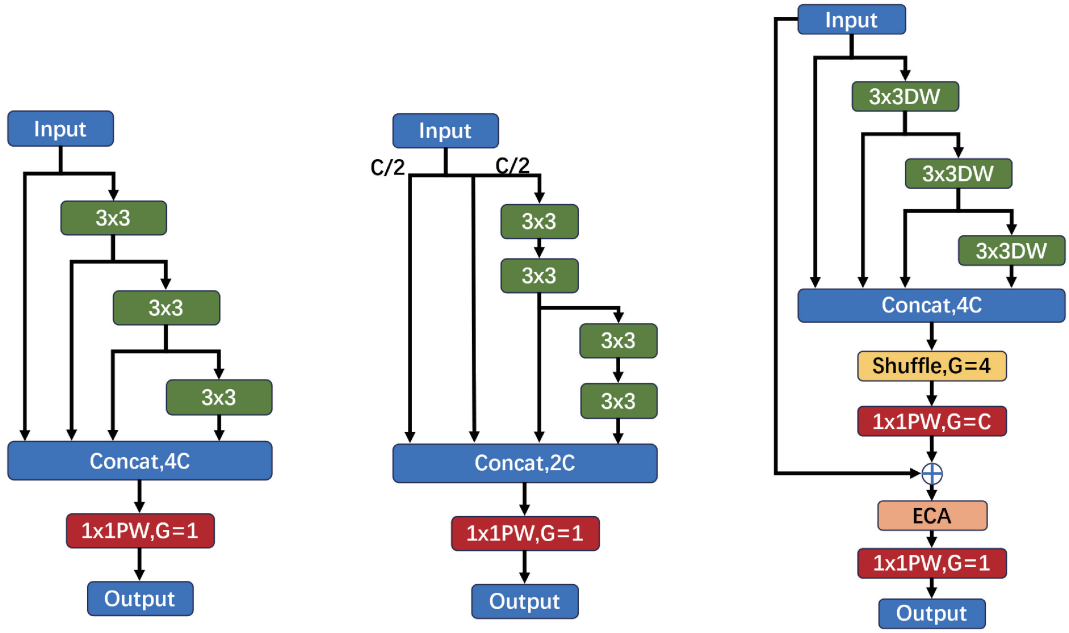
模型结构	输入尺寸	n_1	n_2	n_3	n_4
MFF-Pose-XS	256×256	5	7	9	9
MFF-Pose-S	448×448	5	7	9	9

图 2 MFF-Pose 网络结构

Fig. 2 The whole structure of MFF-Pose

2.1 多层级特征融合模块

卷积操作可以将输入的特征图信息进行归纳概括以获取更为高级且抽象的语义信息,将低级特征图和高级语义图进行适当处理并融合有助于提高模型预测的准确性。VoVNet 提出了将每一个层级卷积之后的结果与此前的结果进行拼接的结构,如图 3(a)所示,该结构将输入的特征图按照不同尺度的感受野进行了特征提取与信息归纳,最后统一卷积运算得到最终的输出结果。该结构相比于连续的 $3*3$ 卷积更好地增强了特征提取的能力,但是传统的 $3*3$ 卷积带来了巨大的计算量。如图 3(b)所示,是 ELAN 当中的模块,该结构借鉴了 VoVNet 网络当中的特征提取融合思路,同时结合了 CSPNet 当中削减计算量的操作方式,在保持了较好的特征提取效果的同时将计算量削减了近一半,但是其使用的 $3*3$ 卷积产生的计算量对于边缘设备依然庞大。此外,前面提到的两种模块都不能进行 shortcut 连接。



(a) VoVNet模块
(a) VoVNet module

(b) ELAN模块
(b) ELAN module

(c) MFF模块
(c) MFF module

图 3　VoVNet、ELAN、MFF 模块结构对比图

Fig. 3　Comparison diagram of VoVNet, ELAN, and multi-stage feature fusion (MFF) block structures

为了提高模型对特定层级进行更强的特征提取能力，同时减轻计算量和参数量，本文提出了一种如图 3(c)所示的轻量级特征提取与融合模块。与 VoVNet 提出的模块相比，MFF 使用深度卷积替换了普通卷积，极大的减少了计算量与参数量。此外由于深度卷积的输出结果是通道无关的，为了专一地对每个特征图自身进行特征提取，不同卷积层级的特征图只需要与属于自己通道的其他特征图进行融合，而没有必要将其他通道的特征图信息混合进来。因此在对每一个卷积层级得到的结果拼接完成后，并没有直接使用逐点卷积进行特征融合和通道调整，而是先对特征图进行了分组数为 4 的通道混洗操作，随后进行了分组数为输入 channel 数的组卷积让不同层级的特征图按照可以学习的权重进行融合。这样，得到的输出是与输入在形状上完全相同，但每一层都进行了增强特征提取和概括的结果，同时，这也允许模块进行 shortcut 连接。

在这之后，添加了通道注意力 ECA 模块(如图 4 所示)，ECA 注意力机制可以注意不同通道之间的关系，为不同的层级赋予不同的权重，重视重要的特征图，忽略次要的特征图，增强模型的表达能力，最后再使用逐点卷积对输出通道进行调整以适应后续模块的输入。与前文提到的两个模块相比，本文提出的模块更加轻量化，同时可以进行 shortcut 连接与通道注意力机制的有效添加，并且可以作为辅助增强模块添加到各种场合当中。

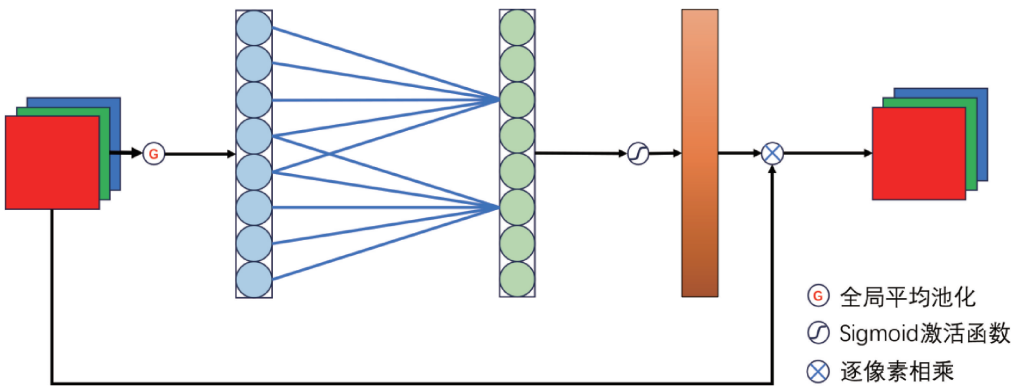


图 4　ECA 注意力机制

Fig. 4　Efficient channel attention block

2.2 高分辨率融合分支设计

高分辨率融合分支通过将特征提取主干网络的不同层级的输出结果。具体的分支结构如图 5 所示,首先通过逐点卷积进行对特征图通道数量调整,随后通过上采样操作使得特征图的尺寸与高分辨率特征图一致,再将两者按对应位置相加的方式进行特征融合。通过这种操作,可以让网络在既保持了高分辨率信息的同时融合了来自低分辨率层级的抽象语义信息,增强了模型的表达能力。考虑到本文提出的 MFF 模块可以有效地提取和增强特征信息,在将特征图与来自下一个阶段的信息进行融合之前,先通过 MFF 模块对相加融合后的特征图进行进一步的特征提取和特征融合,以提高模型对所得的信息的概括和总结能力。

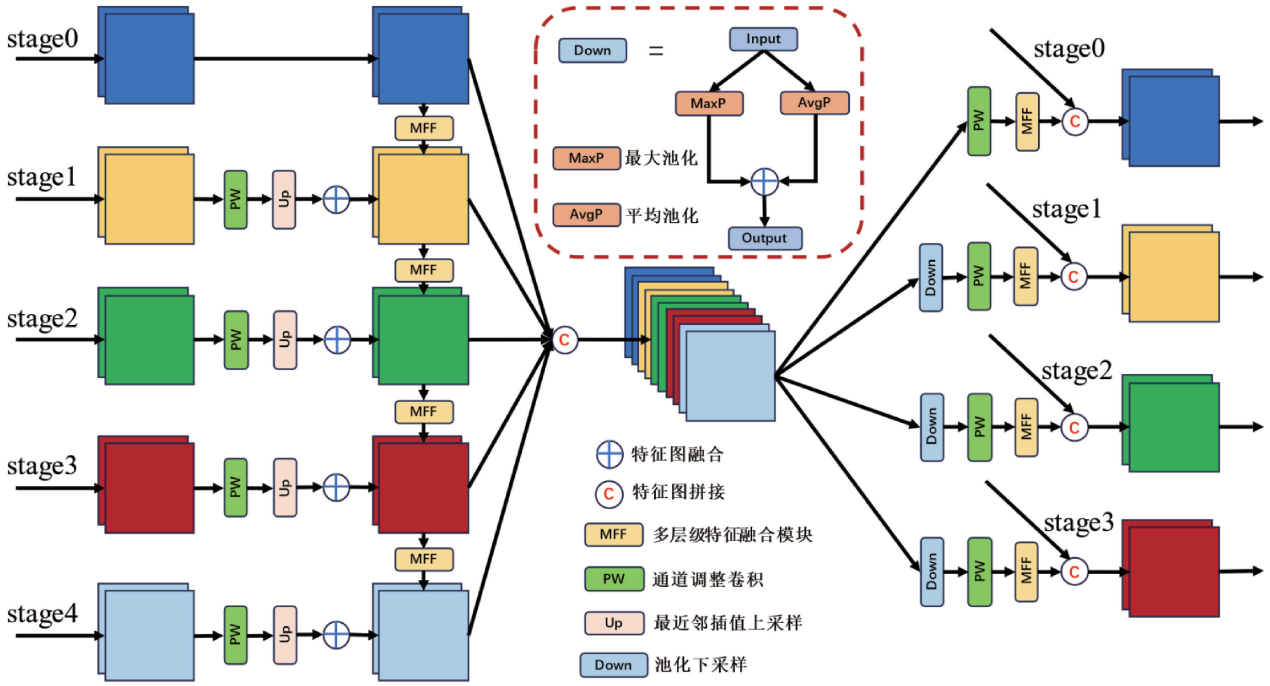


图 5 HRFF 多尺度融合分支

Fig. 5 The structure of high resolution feature fusion (HRFF) block

考虑到连续的卷积操作会导致信息的不断丢失,参考 VoVNet 对特征图的处理方法,本文将来自不同特征提取阶段的经过特征融合后的特征图结果进行拼接,得到了层级丰富的特征图组。随后,将特征图组按照后续上采样操作时不同尺度的特征图大小进行下采样池化和 MFF 特征图增强与通道调整操作,并与原始的来自不同阶段保存的特征图进行拼接,一并送入反卷积层进行上采样处理。通过这种操作,一方面将丰富的特征信息平等的提供给了各个上采样阶段使用,避免了在反卷积过程中的信息丢失,另一方面增加了模型的梯度路径数量,使得模型在进行梯度反向传播时可以获得更强的学习能力。

2.3 AE 分类增强模块

原始 LitePose 网络结构在输出热力图和分类图时仅仅依赖于单个反卷积,而热力图和分类图在期望的输出表现形式上非常不同。人体关键点热力图是对每一个关键点的概率估计,对关键点的定位精度要求越高,其最终呈现的形式越应当趋向于在兴趣区域有类高斯核的概率分布,而在其他区域取得相近且趋向于 0 的结果数值。但对于分类图,其原理要求对于不同的个体,其取值所在的片区取值相距尽可能大,而对于属于同一个个体的不同关节,其取值尽可能相近,这就要求输出结果尽可能地区域化划分,使用单一的卷积难以同时满足这两种需求。

对此,本文提出了针对分类图输出处理的如图 6 所示的分类增强模块结构。考虑到分类图更加注意全局,将输出部分和进行后处理,结合前文设计的 MFF 模块,设计了针对最终 AE 输出的 EAE 模块。该模块分首先将 Heatmap 结果与 AE 分类结果进行拆分,将此前的高分辨率分支进行通道调整和下采样操作转化至和 AE 分类特征图一致的形状,随后将两者进行拼接。为了在进行不同尺度的特征融合的同时将 AE 信息与高分辨率信息进行融合,对拼接后的结果进行通道混洗操作。在混洗操作之后,等分为 2 组,一组直接进入 MFF 模块,另一组先进行 2 倍的卷积下采样,随后再送入 MFF 模块,之后再进行上采样操作,达到和输入相同的尺度。随后将两个部分的结果和原始的 AE 结果进行相加,再与此前输出的热力图部分进行拼接,得到最终的输出结果。

3 实验与分析

3.1 实验设置

本算法的实验环境为:Ubuntu18.04 操作系统,

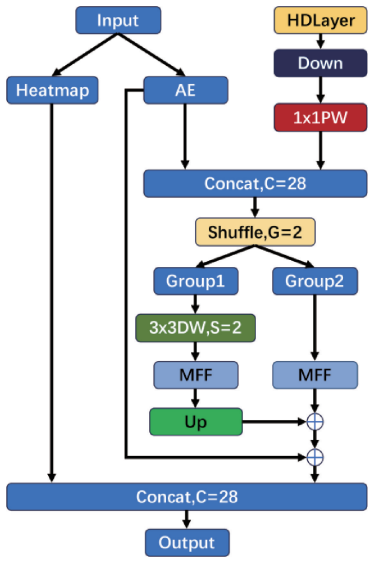


图 6　EAE 分类增强模块

Fig. 6　Enhanced associated embedding (EAE) block

Python3.8.10,Pytorch 2.0.0,CUDA 11.8;其实验硬件环境为:Intel(R) Xeon(R) Platinum 8362,内存 64 GB,2 块 RTX 3090(24 GB)。在模型训练过程中,先保持 LitePose 的原始结构在 CrowdPose 训练集上进行 Supernet 训练,batch size 大小设置为 32,学习率设置为固定值 1×10^{-3},训练 800 轮。随后,按照 LitePose 的原始搜索配置,提取出结构分别为 XS 与 S 两组权重参数。将参数作为预训练参数加载到本文所提出的改进的模型结构当中做后续训练。

对于 CrowdPose 数据集,设置输出关键点个数为 14,在 CrowdPose 训练集上进行训练,batch size 大小设置为 32,初始学习率设置为 1×10^{-3},进行 50 次迭代,随后将学习率削减至 1×10^{-4},进行 150 次迭代,最后将学习率削减至 1×10^{-5},进行 20 次迭代。

对于 MS COCO 数据集,在 MS COCO 训练集上训练,batch size 大小设置为 32,初始学习率设置为 1×10^{-3},进行 50 次迭代,随后将学习率削减至 1×10^{-4},进行 120 次迭代,最后将学习率削减至 1×10^{-5},进行 30 次迭代。所有参数都使用 Adam 优化器进行更新。

3.2　数据集简介及预处理

本文采用 MS COCO 数据集与 CrowdPose 数据集进行训练和验证。

MS COCO 关键点检测数据集包含 64 000 张图像,其中有 27 000 个行人实例,每个人体姿态实例包含 17 种人体关键点。其中,MS COCO 的训练集包含 57 000 张图像,总共有 150 000 个行人实例,验证集包含 5 000 张图像,测试集包含 20 000 张图像。

CrowdPose 数据集中总共包含 20 000 张图像,80 000 个人体姿态实例。每个人体姿态实例包含 14 个关键点。相较于 MS COCO 数据集,CrowdPose 数据集包含了更多人群聚集的场景,使得人体姿态估计任务变得更加艰巨。数据集当中训练集、验证集、测试集的划分比例为 5∶1∶4。

本文使用 CrowdPose 数据集进行预先训练,对输入的图片进行数据增强。数据增强操作包括将输入的图像随机旋转[−30°,30°]、进行[0.75,1.5]的随机尺度变换、进行[−40,40]的随机平移变换和图像翻转。数据增强不改变训练集图像的输入个数,仅在数据输入前对图像与相应的 groundtruth 坐标进行随机变换,让模型可以更加灵活的学习,提高模型对不同场景识别的鲁棒性。

3.3　评价指标

MS COCO 数据集与 CrowdPose 数据集均同样采用关键点相似度(object keypoint similarity,OKS)准确率的平均值(mean average precision, mAP)作为算法性能的评价指标,OKS 的计算公式如公式(1)所示。

$$T_{OKS}=\frac{\sum_i \exp(-d_i^2/2s^2k_i^2)\delta(v_i>0)}{\sum_i \delta(v_i>0)} \tag{1}$$

式中:d_i 表示第 i 个人体姿态关键点标注的位置和远测关键带你位置之间的欧氏距离;s 表示目标尺度,等于该目标在真实图像中的面积平方根;k 表示每一个关键点的衰减系数;v_i 表示仍提关键点在图中是否可见($v_i>0$ 代表能够观察到关键点的位置,$v_i\leqslant0$ 代表不能观察到关键点的位置),最后运算出的每个关键点的相似度的取值介于 0～1。

在模型准确度评估方面,mAP 表示 T_{OKS} 在 0.5～0.95 区间上以 0.05 为间隔每次计算以此 AP 数值,再取所有结果的平均值作为最终结果。分数越高表明该模型检测的效果越好。AP^{50} 表示当 $T_{OKS}=0.5$ 时的检测准确率,AP^{75} 表示当 $T_{OKS}=0.75$ 时的准确率。

在模型参数量和复杂度评估方面,参数量统计模型的实际单数个数,以 M 为单位进行计数,计数越低表明模型的参数量越少。模型的复杂度使用操作数(MACs)作为评价指标,以 G 为单位进行计数,操作数越低表明模型消耗的计算量越小,对边缘设备越友好。

3.4　实验结果及分析

1)在 CrowdPose 测试集上的结果分析

在表 1 中报告了提出的网络在 CrowdPose test 数据集中和不同文献的方法在准确度和参数量上的对比。实验结果如表 1 所示。

本文提出的网络在输入大小为 256×256 的情况下平均精度为 50.7%,相比没有使用知识蒸馏的 LitePose 基准模型提高了 2.3%。和使用了知识蒸馏的模型进行对比,模型也提升了 1.2%,并且相比于 LitePose-XS 仅增加 0.05 M 参数量和 0.38 GMACs 计算量的情况下超过了 HigherHRNet-W16 的 0.3%,同比计算量削减了 10.92 GMACs。与 EffcientHRNet-H-3 相比,提高了 4.5%,计算量削减了 2.72 GMACs。

表 1 不同算法在 CrowdPose test 数据集上的结果

Table 1 Results of different algorithms on the CrowdPose test dataset

方法	输入尺寸	Params/M	MACs/G	AP/%	AP^{50}/%	AP^{75}/%
HigherHRNet-W48[11]	640×640	63.8	154.6	65.9	86.4	70.6
Scaled-HigherHRNet-W16[15]	512×512	7.2	12.5	50.4	78.4	54.5
EffcientHRNet-H_{-3}[28]	416×416	5.3	4.3	46.1	79.3	48.3
LitePose-XS[15](不使用蒸馏)	256×256	1.70	1.20	48.4	—	—
LitePose-XS[15]	256×256	**1.70**	**1.20**	49.5	74.5	51.4
MFF-Pose-XS(本文)	256×256	1.75	1.58	**50.7**	**75.8**	**52.4**
Scaled-HigherHRNet-W24[15]	512×512	14.9	25.3	57.4	83.2	63.2
EffcientHRNet-H_{-1}[28]	480×480	13.0	14.2	56.3	81.3	59.0
LitePose-S[15]	448×448	**2.70**	**5.00**	58.3	81.1	61.8
MFF-Pose-S(本文)	448×448	2.81	6.23	**59.1**	**82.0**	**62.6**

在输入为 448×448 的情况下平均精度为 59.1%,相较于基准 LitePose 模型提高了 0.8%,计算量增加了 1.23 GMACs。相较于 HigherHRNet-W24 提高了 1.7%,计算量削减了 19.07 GMACs,相较于 EffcientHRNet-H-1 提高了 2.8%,计算量削减了 7.97 GMACs。

在 CrowdPose 测试集上的识别结果如图 7 所示。为了直观比对模型的识别结果,将人体关键点评估阈值从 0.1 调整至 0.5,较高的阈值意味着识别的网络对关键点的确信度要更高才能被接受。此时 LitePose-XS 的最终平均精度从 49.6%下滑至 44.9%,下滑明显,而 MFF-Pose-XS 从 50.7%下滑至 49.4%。结果中第一行内容为 LitePose-XS 模型在阈值为 0.5 下的识别结果,第二行为相同阈值下 MFF-Pose-XS 的识别结果。通过对比可以发现,改进前的 LitePose 模型存在关键点未检出(例如最后一组图中男人的膝盖)与关节错误归类的情况(例如中间图中男子的膝盖的同一个部分被同时归类为左右膝盖)。相较而言 MFF-Pose 在这些问题上处理的更好,识别的更加精确。

图 7 人体姿态估计算法 LitePose 和 MFF-Pose 在 CrowdPose 测试集上的识别结果对比

Fig. 7 Comparison of recognition results of LitePose and MFF-Pose on the CrowdPose test set for human pose estimation

2)在 MS COCO val2017 验证集上的结果分析

在表 2 中报告了提出的网络在 MS COCO val2017 验证集中和不同文献的方法在准确度和参数量上的对比。实验结果如表 2 所示。

本文提出的网络在输入大小为 256×256 的情况下平均精度为 41.9%,相比经过知识蒸馏后的 LitePose 基准模型提高了 1.3%,与 EfficientHRNet-H_4 相比提高了 6.2%。

在输入为 448×448 的情况下平均精度为 57.0%,相较于基准 LitePose 模型提高了 0.2%。相较于 EfficientHRNet-H_2 提高了 4.1%,与 Lightweight OpenPose 网络相比,平均精度提高了 14.2%。

在 MS COCO val2017 验证集上的识别结果如图 8 所示。通过采用和 CrowdPose 数据集中一样的评估策略将关键点识别阈值从 0.1 调整至 0.5 后,LitePose-XS 的识别

表 2　不同算法在 MS COCO val2017 数据集上的结果

Table 2　Results of different algorithms on the MS COCO val2017 dataset

方法	输入尺寸	Params/M	MACs/G	AP/%
PersonLab[29]	1 401×1 401	68.7	405.5	66.5
Hourglass[8]	512×512	277.8	206.9	56.6
HigherHRNet-W48[11]	640×640	63.8	155.1	69.9
EfficientHRNet-H_4[28]	384×384	2.80	2.2	35.7
LightPose-XS[15]	256×256	**1.70**	**1.21**	40.6
MFF-Pose-XS(本文)	256×256	1.75	1.58	**41.9**
Lightweight OpenPose[30]	368×368	4.10	9.0	42.8
EfficientHRNet-H_2[28]	448×448	8.30	7.9	52.9
LightPose-S[15]	448×448	**2.70**	**5.0**	56.8
MFF-Pose-S(本文)	448×448	2.81	6.2	**57.0**

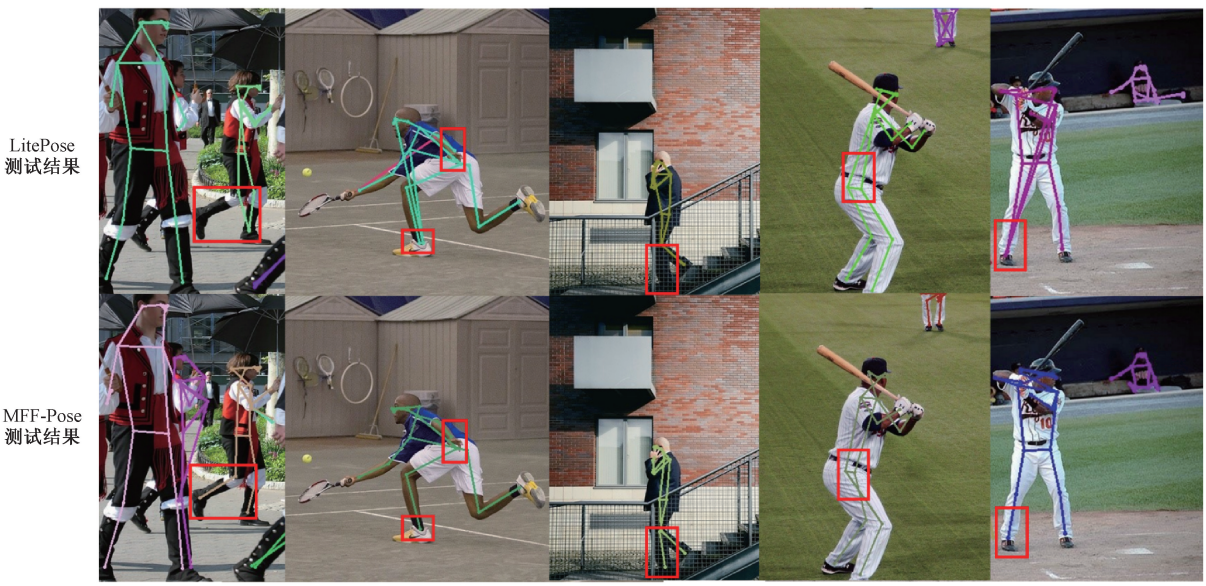


图 8　人体姿态估计算法 LitePose 和 MFF-Pose 在 MS COCO val2017 验证集上的识别结果对比

Fig. 8　Comparison of recognition results of LitePose and MFF-Pose on the MS COCO val2017 set for human pose estimation

结果的平均精度从 40.6%下降至 36.6%，而 MFF-Pose-XS 的结果从 41.9%下降至 39.9%，与 LitePose 相比，MFF-Pose 的稳定性更高。结果中第一行内容为 LitePose-XS 模型在阈值为 0.5 下的识别结果，第二行为相同阈值下 MFF-Pose-XS 的识别结果。与在 CrowdPose 数据集的结果一致，改进前的网络存在大量的误识别结果和漏识别结果，改进后的网络在识别精度上有着显著的提升。

3)消融实验

为了进一步验证本文提出的算法性能，本文在 CrowdPose 数据集上进行消融实验。其实验结果如表 3 所

表 3　算法在 CrowdPose test 数据集上的消融实验

Table 3　Ablation experiments of the algorithm on the CrowdPose test dataset

方法	HDF	MFF	EAE	AP/%	Params/M	MACs/G
LitePose-XS[15]				49.5	1.70	1.20
LitePose-XS[15](无知识蒸馏)				48.4	1.70	1.20
	√			50.4	1.96	2.86
	√	√		50.5	1.74	1.57
	√	√	√	50.7	1.75	1.58

示。其中第一行选择用 LitePose 未经过知识蒸馏操作之后得到的结果 48.4%作为基准。第二行仅添加了高分辨率分支融合,分支与网络主干的连接使用了普通的深度可分离卷积组合,相比于原始结果获得了 2.0%的提升;第三行使用多阶段特征融合模块替换深度可分离卷积组合后,精度继续上升 0.1%,同时参数量和操作数均有下降;第四行添加了 AE 分类增强模块,在增加了 0.01 M 的参数和 0.01 GMACs 的操作数后,识别结果获得了 0.2%的提升,最终的平均精度得到了 50.7%。

4 结 论

为了保持轻量级卷积神经网络参数量与操作数都较低的情况下,提高人体姿态估计网络的识别性能。本文提出了一种基于多层级特征融合与高分辨率图融合的轻量级人体姿态估计网络模型 MFF-Pose。由于多阶段特征融合模块可以有效地提高特征图对信息的挖掘能力同时让通道层级变换也可以进行输入和输出之间的短连接,使得模型可以更好地学习学习不同尺度特征的关系。由于高分辨率融合分支的存在,可以让模型在最后的预测端口能够获得来自不同层级阶段与不同尺度融合的信息,判断的信息更加丰富,提高了模型最终判断的准确度。此外,本文设计提出的两个模块相对于 LitePose-XS 模型只增加了 0.05 M 的参数和 0.38 GMACs 的操作数。未来的工作将所提出的轻量化模块进一步的用于其他领域的特征融合模型当中,可以更加有效地对特征图的信息进行多尺度,多层级的融合。

参考文献

[1] 李一凡,袁龙健,王瑞. 基于 OpenPose 改进的轻量化人体动作识别模型[J]. 电子测量技术,2022,45(1):89-95.
LI Y F, YUAN L J, WANG R. Improved lightweight human action recognition model based on OpenPose[J]. Electronic Measurement Technology, 2022, 45(1): 89-95.

[2] 王鑫,郑晓岩,高焕兵,等. 基于卷积神经网络和多判别特征的跌倒检测算法[J]. 计算机辅助设计与图形学学报, 2023,35(3):452-462.
WANG X, ZHENG X Y, GAO H B, et al. A fall detection algorithm based on convolutional neural network and multi-discriminant feature[J]. Journal of Computer-Aided Design & Computer Graphics, 2023, 35(3): 452-462.

[3] 李志晗,刘银华,谢锐康,等. 基于关节点运动估计的人体行为识别[J]. 电子测量技术, 2022, 45(24): 153-160.
LI ZH H, LIU Y H, XIE R K, et al. Human action recognition based on joint motion estimation[J]. Electronic Measurement Technology, 2022, 45(24): 153-160.

[4] 叶彦斐,胡龙癸,张成龙. 基于改进 YOLOv8n-Pose 的轨道作业人员跨轨安全动作识别[J]. 国外电子测量技术 2024,43(8):181-188.
YE Y F, HU L G, ZHANG CH L. Safety actions recognition of rail workers crossing the track based on improved YOLOv8n-Pose [J]. Foreign Electronic Measurement Technology, 2024, 43(8): 181-188.

[5] 朱周华,侯智杰,田成源,等. 基于改进 YOLOv8-pose 的分心驾驶检测与识别[J]. 电子测量技术, 2024, 47(15):135-143.
ZHU ZH H, HOU ZH J, TIAN CH Y, et al. Distracting driving detection and identification based on an improved YOLOv8-pose [J]. Electronic Measurement Technology, 2024, 47(15): 135-143.

[6] THSHEV A, SZEGEDY C. Human pose estimation via deep neural networks[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014: 1653-1660.

[7] TOMPSON J, JAIN A, LECUN Y, et al. Joint training of a convolutional network and a graphical model for human pose estimation[J]. Advances In neural Information Processing Systems, 2014: 1799-1807,DOI:10.48550/arXiv.1406.2984.

[8] NEWELL A, YANG K Y, DENG J. Stacked hourglass networks for human pose estimation[C]. European Conference on Computer Vision (ECCV), Netherlands: Springer, 2016: 483-499.

[9] 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.

[10] CHEN Y L, WANG Z C, PENG Y X, et al. Cascaded pyramid network for multi-person pose estimation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 7103-7112.

[11] CHENG B W, XIAO B, WANG J D, et al. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation [C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 5386-5395.

[12] WANG J D, SUN K, CHENG T H, et al. Deep high-resolution representation learning for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 43(10): 3349-3364.

[13] YU C Q, XIAO B, GAO C X, et al. Lite-hrnet: A

lightweight high-resolution network[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2021: 10440-10450.

[14] 谢唯嘉,易见兵,曹锋,等.基于特征增强的高分辨率人体姿态估计网络[J].电子测量技术,2024,47(2):131-141.
XIE W J, YI J B, CAO F. High-resolution human pose estimation network based on feature enhancement[J]. Electronic Measurement Technology, 2024, 47(2): 131-141.

[15] WANG Y H, LI M Y, CAI H, et al. Lite pose: Efficient architecture design for 2d human pose estimation[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 13126-13136.

[16] HUANG G, LIU Z, LAURENS V D M, et al. Densely connected convolutional networks[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2017: 4700-4708.

[17] HE K M, ZHANG X Y, REN SH Q, et al. Deep residual learning for image recognition[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.

[18] LEE Y W, HWANG J W, LEE S R, et al. An energy and GPU-computation efficient backbone network for real-time object detection[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019.

[19] WANG C Y, LIAO H Y M, WU Y H, et al. CSPNet: A new backbone that can enhance learning capability of CNN[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020: 390-391.

[20] WANG C Y, LIAO H Y M, YEH I H. Designing network design strategies through gradient path analysis[J]. ArXiv preprint arXiv:2211.04800, 2022.

[21] WANG C Y, BOCHKOVSKIY A, LIAO H Y M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 7464-7475.

[22] WANG C Y, YEH I H, LIAO H Y M. YOLOv9: Learning what you want to learn using programmable gradient information[C]. European Conference on Computer Vision. Springer, Cham, 2025: 1-21.

[23] HU J, SHEN L, SUN G, et al. Squeeze-and-excitation networks[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2018: 7132-7141.

[24] WOO S Y, PARK J C, LEE J Y, et al. CBAM: Convolutional block attention module[C]. European Conference on Computer Vision(ECCV),2018: 3-19.

[25] WANG Q L, WU B G, ZHU P F, et al. ECA-Net: Efficient channel attention for deep convolutional neural networks[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 11534-11542.

[26] HOU Q B, ZHOU D Q, FENG J S. Coordinate attention for efficient mobile network design[C]. the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 13713-13722.

[27] 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.

[28] NEFF C, SHETH A, FURGURSON S, et al. Efficienthrnet: Efficient scaling for lightweight high-resolution multi-person pose estimation[J]. ArXiv preprint arXiv:2007.08090, 2020.

[29] PAPANDREOU G, ZHU T, CHEN L C, et al. Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model[C]. European Conference on Computer Vision(ECCV),2018: 269-286.

[30] CAO Z, SIMON T, WEI S E, et al. Realtime multi-person 2d pose estimation using part affinity fields[C]. IEEE Conference on Computer Vision and Pattern Recognition,2017: 7291-7299.

作者简介

施昕昕,博士,教授,主要研究方向为机器人控制、自抗扰控制。

E-mail: sxx@njit.edu.cn

张昊亮(通信作者),硕士研究生,主要研究方向为图像处理、深度学习、人体姿态估计网络设计。

E-mail: 15751767913@163.com