DOI:10.19651/j.cnki.emt.2314508

基于特征增强的高分辨率人体姿态估计网络*

谢唯嘉 易见兵 曹 锋 李 俊
(江西理工大学信息工程学院 赣州 341000)

摘 要:在轻量级卷积神经网络进行高分辨率人体姿态估计时存在提取特征不充分,针对该问题,提出了一种基于特征增强的高分辨率人体姿态估计网络。首先利用空洞卷积补全操作提取图像特征,以避免特征信息丢失且保持模型参数基本不变;接着利用池化增强模块进行卷积提取特征的选择,以保留重要特征且减轻传统池化模块对提取特征造成的破坏;最后利用加强通道信息交互的深度可分离卷积模块进行特征提取,以保持该模块的参数量较少且能够提高其特征提取能力。在 COCO2017 数据集进行性能测试,本文算法和 DiteHRNet30 算法的 *AR* 值分别为 77.9%和 77.2%;在 MPII 数据集进行性能测试,本文算法和 DiteHRNet30 算法的 *PCKh* 值分别为 32.6%和 31.7%。实验结果表明,本文算法在人体姿态估计精度和算法复杂度之间能够达到较好的平衡。

关键词:人体姿态估计;轻量级网络;高分辨率;空洞卷积;池化;深度可分离卷积
中图分类号:TP391.41 **文献标识码**:A **国家标准学科分类代码**:520.20

High-resolution human pose estimation network based on feature enhancement

Xie Weijia Yi Jianbing Cao Feng Li Jun
(School of Information Engineering, Jiangxi University of Science and Technology, Ganzhou 341000, China)

Abstract: In order to solve the problem of insufficient extracted features in high-resolution human pose estimation using lightweight convolutional neural network, a high-resolution human pose estimation network based on feature enhancement is proposed in this paper. Firstly, the dilated convolution completion operation was used to extract image features to avoid the loss of feature information and basically keep the model parameters unchanged. Then, the pooling enhancement module was used to select the features of convolution extraction, which retained important features and reduced the damage caused by traditional pooling module on extracted features. Finally, the depthwise separable convolution module that strengthens the channel information interaction was used for feature extraction, so as to keep the number of parameters of the module small and improve its feature extraction ability. The performance of the proposed algorithm and DiteHRNet-30 algorithm were tested on the COCO2017 dataset. The *AR* values of the proposed algorithm and DiteHRNet-30 algorithm are 77.9% and 77.2%, respectively. The performance of the proposed algorithm and DiteHRNet-30 algorithm are tested on the MPII dataset. The *PCKh* values of the proposed algorithm and DiteHRNet-30 algorithm are 32.6% and 31.7%, respectively. Experimental results show that the proposed algorithm can achieve a good balance between the accuracy of human pose estimation and the complexity of the algorithm.

Keywords: human pose estimation; lightweight network; high resolution; dilated convolution; pooling; depth separable convolution

0 引 言

人体姿态估计是计算机理解人体动作和行为的关键环节,其对准确性和实时性要求较高,在人体行为识别[1-2],运动捕捉[3]和交通[4]等领域都具有广泛的应用前景。在人体姿态估计任务中,获得高分辨率特征图至关重要,但会大幅

收稿日期:2023-09-01
*基金项目:国家自然科学基金(62066018)、江西省自然科学基金(20181BAB202004)、江西省教育厅科技项目(GJJ210828,GJJ200818,GJJ180482)、江西省赣州市科技计划项目、江西省研究生创新专项(YC2022-S640)资助

度增加模型复杂度和计算量，导致模型的实时性较差。

基于深度学习的二维人体姿态估计方法[5]可分为单人人体姿态估计和多人人体姿态估计。单人人体姿态估计方法 DeepPose[6]是一种基于关键点坐标回归的方法，但是其训练的模型泛化性较差。Tompson 等[7]首次将热力图回归方法引入人体姿态估计领域中，Huang 等[8]针对热力图回归中存在的统计误差，提出了无偏数据处理的人体姿态估计方法。2016 年，Newell 等[9]提出的沙漏网络是二维人体姿态估计领域中的里程碑，在此基础上，Sun 等[10]基于沙漏网络提出了高分辨率特征表示网络 HRNet，后续工作几乎都以其作为基干网络。在多人姿态估计方法中，旷世科技提出的级联金字塔方法[11]使得姿态估计模型对具有复杂背景或遮挡情况下的人体关键点检测准确率进一步提高。Zhang 等[12]结合难例挖掘的关键点提取方法，提高了网络柔性关节点的定位精度。Dong 等[13]结合实际复杂煤矿场景进行模型优化以达到较高准确度。随着人体姿态估计技术的发展[14-19]，为了实现模型的轻量级部署以达到实际应用的目的，研究者开始聚焦于人体关键点检测的轻量级网络结构设计。为了降低 HRNet 网络复杂度，Small-HRNet[20]通过降低网络的宽度和深度来简化网络结构，但这种方法会导致关键点检测性能的下降。轻量级高分辨率网络 Lite-HRNet[21]利用一个高效的卷积神经网络(convolutional neural network，CNN)模块替换 Small-HRNet 中的残差模块，从而获得令人满意的性能，然而这种方法没有在高分辨率网络中进行验证，其操作可能在网络中的不同位置和对不同大小的输入产生不同的影响[22]。而在轻量级高分辨率网络中，依赖于输入数据的组件单元比不依赖于输入数据的组件单元表现得更有效，Dite-HRNet[23]解决了以往高分辨率网络与输入无关且缺乏不同特征信息的问题。当前的高分辨率网络[8,10,23-25]主要依赖于并行分支中深度堆叠的卷积层提取多尺度特征来建立空间依赖，但轻量级网络的参数量会受到所堆叠的卷积层宽度和深度限制影响，导致该类方法提取的高分辨率特征细节不足，从而影响模型整体的推理性能。

为了解决轻量级卷积神经网络 Dite-HRNet[23]进行高分辨率人体姿态估计时存在提取特征不充分问题，本文提出了一种轻量级高分辨率网络(LH-HRNet)，该网络以 Dite-HRNet[23]为基干网络，在网络中设计了几种轻量级卷积网络模块来提取高分辨率图像中的特征信息，以平衡网络参数量和其检测精度之间的矛盾。Dite-HRNet 算法[23]通过堆叠动态拆分卷积模块和自适应上下文建模模块以达到提高模型预测精度的目的，而本文是通过优化网络中模块的性能以达到充分提取特征的目的。本文首先将空洞卷积核补全(dilated convolution completion，DCC)操作加入到网络结构中的高分辨率特征提取阶段，以避免特征信息丢失；接着引入轻量级池化增强模块(pooling enhancement module，PEM)来保留重要特征且减轻传统池化模块对提取特征造成的破坏；最后本文提出了一种加强通道信息交互的深度可分离卷积模块(improving depthwise separable convolution，IDSC)，该模块可以在保持较低参数量的前提下提高其特征提取能力。

1　相关工作

1.1　高效的 CNN 模块

高效的 CNN 模块已被广泛应用于许多 CNN 架构中[26-28]，其目的是在有限的计算成本下最大化模型容量以提高模型预测的准确度。MobileNetV2[27]采用 Depth-Wise 卷积搭配 Point-Wise 卷积的方式来提取特征，可以成倍的减少卷积层的时间和空间复杂度。ShuffleNetV2[15]中的 Shuffle 模块只对特征中的一半通道特征进行卷积，之后按通道进行特征通道拆分和特征通道洗牌操作[26]，增强了特征在不同通道之间的信息交换。Sand-glass 模块[28]翻转了反向传播模块的结构，在没有任何额外计算成本的情况下减少了特征信息损失。

1.2　空洞卷积

空洞卷积最初是在小波分解算法中发展起来的，其主要思想是在卷积核像素之间插入“孔”(零)，以提高特征的分辨率，从而实现深度卷积神经网络中的密集特征提取。在分割任务中，Yu 等[29]串行叠加使用具有不同空洞率的空洞卷积来实现上下文特征聚合，而 Chen 等[30]设计了一个“阿特劳斯空间金字塔池化”方案，通过并行放置多个扩展卷积层来捕获多尺度特征对象和上下文信息。

2　轻量级高分辨率人体姿态估计网络设计

针对轻量级高分辨率人体姿态估计网络中在利用卷积神经网络提取特征不充分问题，本文提出了一种基于特征增强的高分辨率人体姿态估计网络。该网络的空洞卷积补全操作既能避免空洞卷积提取特征不充分问题，又能保留空洞卷积原有的感受野；该网络的池化增强模块对空洞卷积提取的高分辨率特征进行选择，保留重要特征信息且可以减轻传统池化对提取特征造成的破坏；最后各分支特征聚合时都保持一个高分辨率特征表示，并采用加强通道信息交互的深度可分离卷积模块进一步提升模型高分辨率特征的表达能力，以达到为后续人体姿态估计提供更具质量的高分辨特征。

本文的网络结构如图 1 所示，该网络整体是一个四阶段的卷积神经网络，主要由 1 个高分辨率特征主分支和 3 个由高到低分辨率特征分支组成，这些分支在每个新阶段开始时逐个并行地添加到网络中。与之前添加的分支相比，每个新添加的分支特征大小都只有之前一半的分辨率和两倍的通道数。

本文网络在 4 个阶段的详细信息如表 1 所示，第一阶段之前也被认为是 Stem，由于考虑到该阶段中的特征图分辨率较高，对整体网络性能提升是最大的，所以在这个阶段

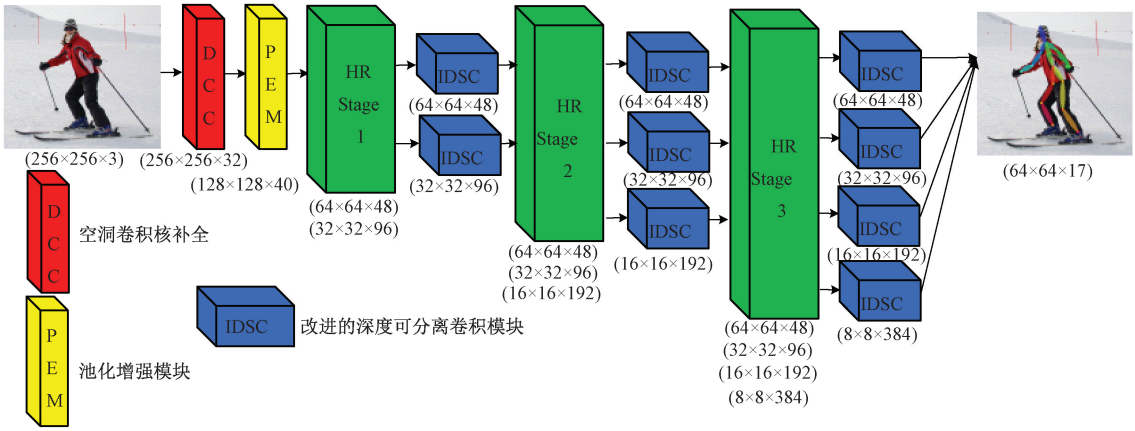


图 1 人体姿态估计网络结构

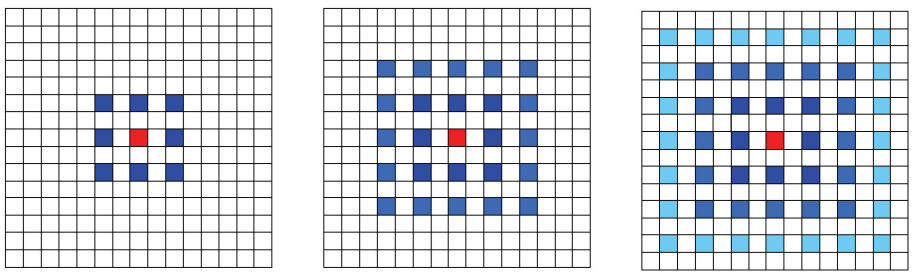
表 1 LH-HRNet 的网络构成模块

层级	输出大小 (输入为 256×256)	操作类型	模块个数	LH-HRNet 层级个数	
				LH-HRNet-18	LH-HRNet-30
Stem	64×64	DCC	1	1	1
		PEM block	1		
Stage 1	64×64	IDSC block	2	2	3
		Multi-scale fusion block	1		
Stage 2	64×64	IDSC block	3	4	8
		Multi-scale fusion block	1		
Stage 3	64×64	IDSC block	4	2	3
		Multi-scale fusion block	1		

采用了空洞卷积核补全操作和池化增强模块来提高特征的表达能力,后续每个阶段由一系列交叉分辨率模块和一个跨所有分支交换信息的多尺度融合模块组成,加强通道信息交互的深度可分离卷积模块在多次尺度特征融合中能够提取有效的分辨率特征信息。

2.1 空洞卷积核补全操作

空洞卷积通过在卷积内核中插入“孔”来提高卷积核对网络特征的感受野,且能够在保持与传统卷积同样参数量的前提下,增强提取特征的表达能力。然而当前空洞卷积核中存在一个固有问题,该问题定义为“网格化影响”,如图 2 所示,由于卷积核两个像素之间填充了零,因此该卷积核的接受域仅覆盖带有棋盘图案的区域,导致只有卷积核的非零值位置对输入进行采样,丢失了一些邻近信息[31]。



(a) 一层空洞卷积结果 (b) 二层空洞卷积结果 (c) 三层空洞卷积结果

图 2 网格化效应

图 2 从左到右:以计算中心像素点(红色标记)为空洞卷积核中心,像素点(蓝色标记)依次通过三层空洞卷积后卷积核所得到的接受域,空洞卷积的卷积核大小为 3×3,空洞率 $r=1$。空洞卷积只能以一种棋盘式的方式进行特征取样,这样会丢失很大一部分特征信息(当 $r=1$ 时,经过三层空洞卷积后至少丢失 70%特征信息)。因此,采用空洞卷积进行堆叠或利用空洞卷积进行下采样操作,这将导致提取的特征信息变得稀疏,不利于网络的整体学习,

具体包括:1) 局部信息部分缺失;2) 破坏了特征信息在浅层网络和深层网络之间的相关性。针对传统空洞卷积由于网格化效应导致提取的特征信息被丢失问题,本文提出了一种空洞卷积补全操作方法,具体操作如图 3 所示。

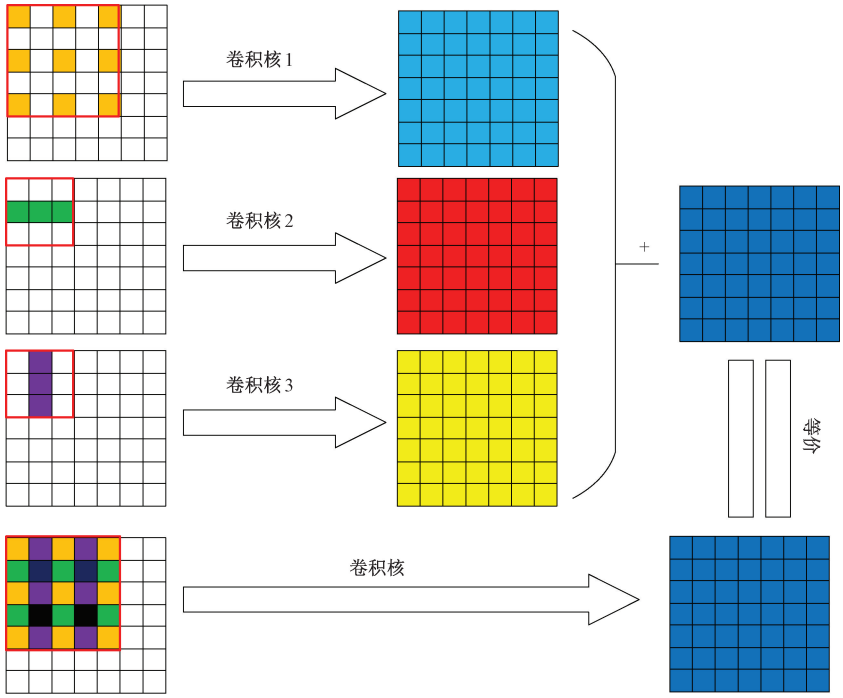


图 3　空洞卷积核补全操作

卷积操作具有如下特点:如果几个大小兼容的二维卷积核以相同的步长对相同的输入进行卷积操作,产生具有相同分辨率的输出,将上述不同卷积操作的输出进行相加,其输出结果和一个等效的卷积核输出相同,如图 3 所示。假设空洞率 $r = 1$ 的空洞卷积、卷积核大小分别为 1×3 和 3×1 的卷积输入相同,通过对输入进行填充和改变卷积的步长,使这 3 个卷积的输出大小相同,因此具有可加性。可加性适用于二维卷积,即在相同的输入情况下,不同卷积核输出进行相加可等价于卷积核相加后再对数据处理的作用效果[32],如式(1)。

$$\boldsymbol{I} \times K^{(1)} + I \times K^{(2)} = I \times (K^{(1)} \oplus K^{(2)}) \tag{1}$$

其中,$\boldsymbol{I}$ 是一个输入矩阵,$K^{(1)}$ 和 $K^{(2)}$ 是两个卷积核大小兼容的二维卷积核,而⊕表示卷积核参数在相应位置上的元素进行相加,但具体操作应注意对不同的输入数据可能需要适当的修剪和填充。根据二维卷积核的可加性,三个卷积的输出相加等价于一个全新的二维卷积核,如图 3 所示,可以直观的看到传统空洞卷积核 1 中所插入的“孔”能正好被卷积核 2 和卷积核 3 进行填充。因此最后等价出的卷积核将特征信息都提取到,不会因为空洞卷积的卷积核缺陷而丢失特征信息。

2.2　池化增强模块

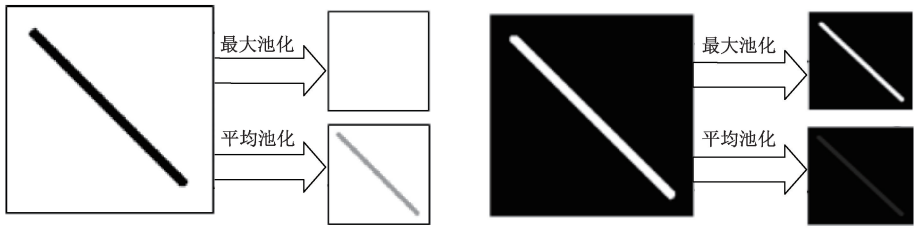
卷积层对输入进行特征映射主要的局限性在于卷积层只能被精确记录到提取的特征在输入层位置。这意味着对输入图像的小幅移动会导致图像中卷积层提取的特征偏移,进而会导致卷积层不同的特征映射。而这些输入图像的小幅度移动在数据增强操作中经常涉及:图像重新裁剪,旋转,移动以及其它数据增强操作。对于以上的特征偏移,池化操作能够很好的进行解决。传统深度学习理论认为卷积操作是特征提取,而池化操作是对提取的特征进行选择。在较深的模型结构中,没有池化操作的模型容易过拟合,因为卷积提取出来的原始特征多且比较复杂,而池化就是对提取出来的特征进行选择,即保留重要特征丢弃无效特征。常规的池化操作有最大池化和平均池化,平均池化可以保留图像中更多的背景信息,而最大池化可以保留更多的纹理信息。

但是最大池化操作和平均池化操作都有各自的缺点,最大池化操作只考虑池化区域内的最大像素值而忽略其它像素值,会出现如图 4(a)所示的情况,即经过最大池化后图像中具有区分性的特征消失;平均池化操作计算的是池化区域内所有元素的平均值,会降低池化后新特征图的对比度,在极端情况下,如果有许多零元素,特征图的特性将大大降低,如图 4(b)所示。

由于模型处理的图像是多种多样的,最大池化和平均池化的缺陷导致将池化模块应用在 CNN 上的算法产生负面影响。

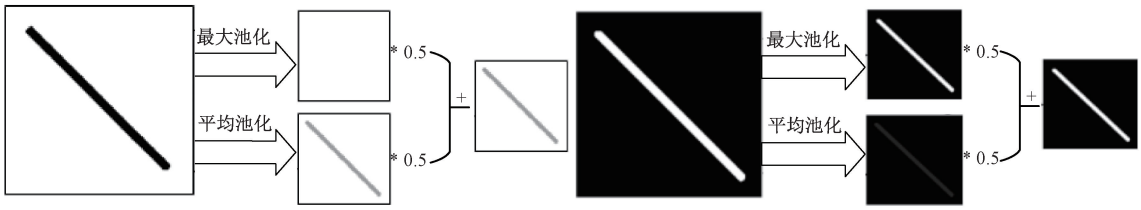
考虑到两种池化操作可相互中和其自身产生的负面影响,如图 5 所示,因此本文提出一种池化增强模块以减少常规池化由于其自身缺陷所带来的影响,如图 6 所示。

考虑到不同池化操作对特征提取时保留的特征不同并能相互减轻其自身缺陷带来的影响,故使用两条不同通



(a) 最大池化的缺陷 (b) 平均池化的缺陷

图 4 最大池化和平均池化的缺陷



(a) 平均池化中和最大池化的负面影响 (b) 最大池化中和平均

图 5 最大池化和平均池化相互中和负面影响

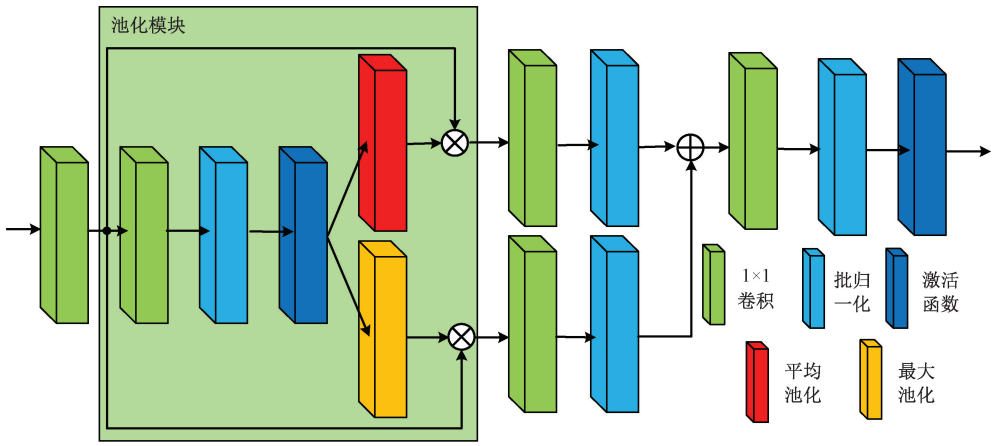


图 6 池化增强模块

路进行不同池化操作。图 6 中的池化模块之前的 1×1 卷积可以保持特征尺度不变的前提下大幅增加网络的非线性特性,保证在网络结构很深的情况下其反向传播的梯度不会消失。而图 6 中的池化模块之后的一系列 1×1 卷积起到了对不同池化操作所得到的通道信息进行跨通道聚合的作用,达到用较少的参数和计算量对特征图像进行通道融合和升维的目的。

考虑到池化操作对卷积所提取出的深层特征是有破坏性的,故池化增强模块只能应用在算法网络中的浅层特征提取中。因为在浅层特征图中,特征的空间相关特性比较明显,可以使用池化模块;而在深层特征图中,特征经过多层卷积操作,其空间相关特性不太明显。

2.3 通道增强深度可分离卷积

深度可分离卷积包括逐深度卷积操作和逐点卷积操作。输入先经过逐深度卷积模块后得到的特征图通道数与输入层的通道数相同,无法进行特征图通道数的扩展。因此需要将这些特征图进行逐点卷积,生成新的特征图。逐点卷积的运算与常规卷积运算非常相似,它的卷积核尺寸为 $M\times1\times1$,M 为上层逐深度卷积模块的通道数,输出的通道数可扩展为 N。

假设输入特征图的大小为 $M\times H_{in}\times W_{in}$,输出特征图的大小为 $N\times H_{out}\times W_{out}$,逐深度卷积操作中卷积核大小为 $N\times K\times K$,逐点卷积尺寸为 $M\times1\times1$,且一共有 N 个逐点卷积滤波器。逐深度卷积,逐点卷积和整体深度可分离卷积的计算复杂度公式分别表示为式(2)~(4)。

$$Depthwise = M\times H_{in}\times W_{in}\times K^2 \tag{2}$$

$$Pointwise = N\times M\times H_{out}\times W_{out} \tag{3}$$

$$Total = M\times H_{in}\times W_{in}\times K^2 + N\times M\times H_{out}\times W_{out} \tag{4}$$

而标准卷积的计算复杂度公式表示为式(5)。

$$Conv = M\times H_{in}\times W_{in}\times K^2\times N \tag{5}$$

从上述推导出的深度可分离卷积与标准卷积的计算

复杂度公式可得，在输入与输出特征图的大小相同情况下（$H_{in}=H_{out}=H$ 和 $W_{in}=W_{out}=W$），其计算复杂度的比值如式(6)。

$$\frac{Total}{Conv}=\frac{M\times H\times W(K^2+N)}{M\times H\times W\times K^2\times N}=\frac{K^2+N}{K^2\times N}\approx\frac{1}{\mathrm{N}} \tag{6}$$

由式(6)可知，深度可分离卷积操作在同等情况下约为标准卷积操作计算量的 $1/N$（N 为输出特征图通道数）。

深度可分离模块的卷积操作对输入层的每个通道独立进行卷积运算，没有有效利用不同通道在相同空间位置上的特征信息。为了解决深度可分离模块在通道方向上特征信息缺失问题，本文提出了一种通道增强深度可分离卷积模块，如图 7 所示。该模块与深度可分离卷积相比，可以在保持较低参数量的前提下大幅提升特征信息在不同通道之间的交互信息以提升模型整体的精度。

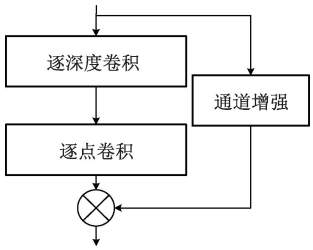


图 7　深度可分离卷积改进模块结构

在深度可分离卷积改进模块中，首先在一条通路中进行深度可分离卷积操作，同时另一条通路利用特征通道信息增强模块对输入进一步提取通道特征信息，最后将提取的通道特征信息融入深度可分离卷积的输出中，得到通道特征增强后的特征图，详细结构如图 8 所示，该结构可以解决深度可分离卷积中由于分组卷积后各个通道的特征信息独立而无法利用不同通道在相同空间位置上的特征信息问题，整体模块定义如式(7)。

$$Y=PWConv(DWConv(X))\times CAM(X) \tag{7}$$

其中，X 和 Y 分别表示的是深度可分离模块的输入和输出，$DWConv$ 表示深度可分离逐深度卷积，$PWConv$ 是整合通道上的特征信息逐点卷积，通道信息增强模块 CAM 表示进一步整合丢失的通道信息。

通道信息增强模块对输入特征图在通道方向进行特征信息提取，之后再将其融入输出特征图上，其相对深度可分离卷积模块能够在小幅度增加推理内存占用的情况下，有效的利用不同通道在相同空间位置上的特征信息。

3　实验与分析

3.1　实验设置

本算法的实验软件环境为：Ubuntu16.04 操作系统，Python3.8.12，PyTorch 1.7.1；其实验硬件环境为：CPU 为 Intel Xeon(R) E5-2630v4，内存 32 G，4 块 Nvidia Tesla P100。在模型训练过程中，Batchsize 设置为 32，所有模型

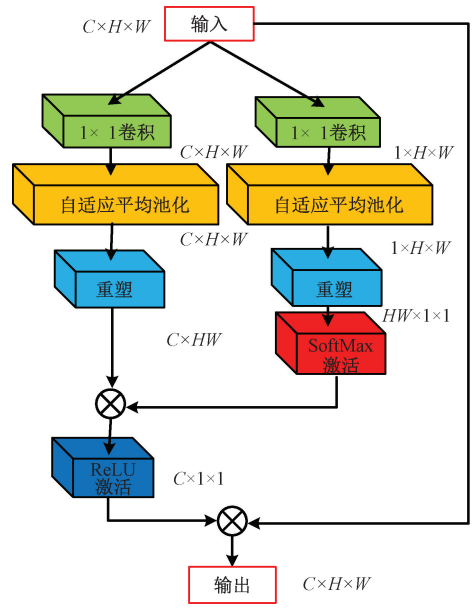


图 8　通道信息增强模块结构

参数都由 Adam 优化器进行更新，其初始学习率设置为 0.002，后续随着模型训练过程进行 Warm Up 操作，最后学习率降至 0.000 02。

3.2　数据集简介及预处理

本文采用 COCO2017[33]、MPII[34] 数据集进行训练和验证。COCO2017 数据集的训练集 train2017 包含 57 000 张图片和 150 000 个人体姿态实例，验证集 val2017 包含 5 000 张图片，每个人体实例有 17 个关键点，分别是鼻子、右眼、左眼、右耳、左耳、右肩、左肩、右肘、左肘、右手腕、左手腕、右臀、左臀、右膝盖、左膝盖、右脚踝、左脚踝。MPII 数据集包含约 25 000 张图片和超过 40 000 个人体姿态实例，其中 12 000 个人体姿态实例用于验证，其它实例用于训练。每个人体实例有 16 个关键点，分别是头顶、上颈部、右肩、左肩、右肘、左肘、右手腕、左手腕、右臀、左臀、盆骨、胸部、右膝盖、左膝盖、右脚踝、左脚踝。

由于 COCO2017 数据集中原始图片的尺寸各不相同，所以需要对原始图片进行数据预处理，然后再进行模型训练。在数据预处理操作中，先将所有人体姿态检测框扩展到固定长宽比为 4∶3，然后以人体的髋部为中心进行裁剪，把图片的尺寸大小重新裁剪为 256×192 和 384×288，之后对训练图片进行一系列的数据增强操作，包括以 30° 的随机旋转因子进行随机翻转，以及以 0.25 的随机尺度缩放因子进行随机缩放。MPII 数据集图片的尺寸大小则调整为 256×256。

模型的验证过程采用两阶段的自顶向下的检测方法，即首先生成人体的检测框，然后再预测人体的关键点。人体检测框由简单基线[35]提供的 COCO 数据集人体检测器进行预测，而 MPII 数据集使用标准测试提供的人体检测框。所生成的热力图是通过二维高斯分布进行估计的，然

后对原始图片和翻转图片进行平均,热力图中概率值的位置再从最高概率值到第二高概率值的方向上调整 1/4 的偏移,以获得最终的关键点位置。

3.3 评价指标

在 COCO2017 数据集中为验证算法模型的性能,采用 MS COCO 官方给定的基于关键点相似度(object keypoint similarity, OKS)准确率的平均值(mean average precison, mAP)作为算法性能评价指标,OKS 的准确率可用式(8)计算得到。

$$T_{OKS}=\frac{\sum_i \exp(-d_i^2/2s^2k_i^2)\delta(v_i>0)}{\sum_i \delta(v_i>0)} \tag{8}$$

式中:d_i 表示第 i 个关键点标注位置和预测关键点位置之间的欧式距离;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_i 表示每个关键点控制的衰减常数;v_i 表示能否观察到真实的人体关键点($v_i>0$ 代表能够观察到关键点位置,$v_i\leqslant 0$ 代表不能观察到关键点位置),最后运算出的每个关键点相似度的值域为 [0,1]。

mAP 表示 T_{OKS} 在 0.5~0.95 区间上每隔 0.05 计算一次 AP 值,再取所有结果的平均值作为最终结果,分数越高,表明该模型在检测和定位图像中的性能越好。AP^{50} 表示当 $T_{OKS}=0.5$ 时的检测准确率,AP^{75} 表示当 $T_{OKS}=0.75$ 时的检测准确率,AP^M 表示目标框像素面积在 32^2~96^2 之间的检测准确率,AP^L 表示目标框像素面积大于 96^2 的检测准确率,AR 表示 T_{OKS} 在 0.50~0.95 区间上每隔 0.05 计算一次召回率,再取所有结果的平均值作为最终结果。

在 MPII 数据集中为验证算法模型的性能,采用模型正确估计出的关键点比例(percentage of correct keypoints, PCK)作为算法的性能评价指标,计算检测的关键点与其对应的真实位置间的归一化距离小于设定阈值的比例。归一化距离是指关键点预测值与人工标注值之间的欧式距离进行人体尺度因子的归一化,而 MPII 数据集是以头部矩形框的左上点与右下点的欧式距离作为尺度因子,使用上述尺度因子的姿态估计指标也称为 $PCKh$,其中 $PCKh@0.1$ 表示设定的阈值为 0.1。

PCK 的数学表达式如式(9)所示。

$$PCK_{mean}^{k}=\frac{\sum_p\sum_i \delta(\frac{d_{pi}}{d_p^{def}}\leqslant T_k)}{\sum_p\sum_i 1} \tag{9}$$

其中,i 表示关键点序号,k 表示阈值的序号,p 表示检测出的人序号,d_{pi} 表示第 p 个人的第 i 个关键点预测值与人工标注值之间的欧式距离,d_p^{def} 表示第 p 个人的尺度因子,T 表示人工设定的阈值,最后所得到模型正确估计出的关键点比例 PCK 的值域为[0,1]。

3.4 实验结果及分析

为了验证算法在进行人体姿态估计时的性能,本文特选取了人体关键点检测领域中多个优秀的算法 UDP[8],Hourglass[9],HRNet[10],CPN[11],ShuffleNetV2[15],Small HRNet[20],Lite-HRNet[21],Dite-HRNet[23],MobileNetV2[27],SimpleBaseline[35],EfficientPoseI[36] 与之进行对比。本文算法与其它人体姿态估计算法在 COCO val2017 与 MPII 两个数据集的实验结果分别在表 2 和 3 中,其中参数和 $GFLOPs$ 这两个性能指标都只是针对人体姿态估计网络,不包括模型对人体整体的检测和对检测出的关键点进行分组操作。

1)在 COCO2017 验证集上的结果分析

本文将 LH-HRNet-18 和 LH-HRNet-30 在 COCO2017 验证集上的实验结果与其它人体姿态估计算法进行比较,包括重量级网络结构和轻量级网络结构,如表 2 所示。本文分别采用输入图片为 256×192 和 384×288 对算法进行测试,从整体上看本文算法比其它算法在精度和模型复杂度之间取得了更好的平衡。与现有的轻量级网络结构相比,本文提出的 LH-HRNet-30 算法在输入图片大小为 384×288 时 AP 值最高,到达了 72.8% 的精度。LH-HRNet-18 在输入图片大小为 256×192 时,达到了 8 阶段的 Hourglass[9] 重量级网络的精度且参数量和 $GFLOPs$ 少了一个数量级。特别是在输入图片大小为 256×192 时,本文提出的 LH-HRNet-18 算法比 Small HRNet[20] 轻量级网络在 AP 指标上高出 11%,且 $GFLOPs$ 仅为 Small HRNet 的 60%;当输入图片大小为 384×288 时,本文算法 LH-HRNet-18 比 Small HRNet[20] 的 AP 高出 15%,$GFLOPs$ 仅为 Small HRNet 的 58%。总之,与重量级网络 UDP[8],Hourglass[9],HRNet[10],CPN[11],SimpleBaseline[35] 相比,本文提出的 LH-HRNet 网络以更小的模型尺寸和更低的计算复杂度实现了相当甚至更高的人体关键点检测精度;与轻量级网络 ShuffleNetV2[15],Small HRNet[20],Lite-HRNet[21],Dite-HRNet[23],MobileNetV2[27] 相比,本文提出的 LH-HRNet 网络以较小的计算复杂度增长得到人体关键点检测精度较大的提升。

2)在 MPII 验证集上的结果分析

本文将 LH-HRNet-18 和 LH-HRNet-30 在 MPII 验证数据集上的实验结果与其它人体姿态估计算法的结果进行比较,如表 3 所示(表 3 中的结果都是基于输入图片大小为 256×256)。LH-HRNet-30 模型与 Dite-HRNet-30 模型相比,在 $PCKh@0.1$ 评价指标提高了 0.9%,在 $PCKh$ 评价指标提高了 0.5%。LH-HRNet-18 模型与 Dite-HRNet-18 模型相比,在 $PCKh@0.1$ 评价指标提高了 0.8%,在 $PCKh$ 评价指标提高了 0.4%。LH-HRNet-18 在参数量,模型复杂度,$PCKh$ 和 $PCKh@0.1$ 等性能指标上都优于 ShuffleNetV2[15],MobileNetV2[27]。本文的 LH-HRNet-18 模型与 Dite-HRNet-30 模型的运算复杂度基本相当,但本文的模型参数量少且其 $PCKh@0.1$ 值比 Dite-HRNet-30 模型提升了 0.2%。与轻量级网络 ShuffleNetV2[15],

表 2　不同算法在 COCO val2017 数据集上的结果

模型	预训练	输入大小	参数/M	*GFLOPs*	*AP*	AP^{50}	AP^{75}	AP^{M}	AP^{L}	*AR*
重量级网络										
UDP[8]	Y	256×192	28.7	7.1	75.2	92.4	82.9	72.0	80.8	80.4
Hourglass[9]	N	256×192	25.1	14.3	66.9	—	—	—	—	—
HRNet[10]	N	256×192	28.5	7.1	73.4	89.5	80.7	70.2	80.1	78.9
CPN[11]	Y	256×192	27.0	6.2	68.6	—	—	—	—	—
SimpleBaseline[35]	Y	256×192	34.0	8.9	70.4	88.6	78.3	67.1	77.2	76.3
轻量级网络										
ShuffleNetV2[15]	N	256×192	7.6	1.2	59.9	85.4	66.3	56.6	66.2	66.4
Small HRNet[20]	N	256×192	1.3	0.5	55.2	83.7	62.4	52.3	61.0	62.1
MobileNetV2[27]	N	256×192	9.6	1.4	64.6	87.4	72.3	61.1	71.2	70.7
Lite-HRNet-18[21]	N	256×192	1.1	0.2	64.8	86.7	73.0	62.1	70.5	71.2
Lite-HRNet-30[21]	N	256×192	1.8	0.3	67.2	88.0	75.0	64.3	73.1	73.3
Dite-HRNet-18[23]	N	256×192	1.1	0.2	65.9	87.3	74.0	63.2	71.6	72.1
Dite-HRNet-30[23]	N	256×192	1.8	0.3	68.3	88.2	76.2	65.5	74.1	74.2
LH-HRNet-18(ours)	N	256×192	1.6	0.3	66.9	87.5	74.9	64.3	72.5	73.0
LH-HRNet-30(ours)	N	256×192	2.6	0.4	69.2	88.6	77.2	66.5	75.1	75.2
ShuffleNetV2[15]	N	384×288	7.6	2.8	63.6	86.5	70.5	59.5	70.7	69.7
Small HRNet[20]	N	384×288	1.3	1.2	56.0	83.8	63.0	52.4	62.6	62.6
MobileNetV2[27]	N	384×288	9.6	3.3	67.3	87.9	74.3	62.8	74.7	72.9
Lite-HRNet-18[21]	N	384×288	1.1	0.4	67.6	87.8	75.0	64.5	73.7	73.7
Lite-HRNet-30[21]	N	384×288	1.8	0.7	70.4	88.7	77.7	67.5	76.3	76.2
Dite-HRNet-18[23]	N	384×288	1.1	0.4	69.0	88.0	76.0	65.5	75.5	75.0
Dite-HRNet-30[23]	N	384×288	1.8	0.7	71.5	88.9	78.2	68.2	77.7	77.2
LH-HRNet-18(ours)	N	384×288	1.6	0.7	70.3	88.4	77.5	66.8	76.8	75.8
LH-HRNet-30(ours)	N	384×288	2.6	1	72.8	89.3	79.5	69.5	78.9	77.9

注：预训练＝在 ImageNet 分类任务上预训练骨干网络结构。指标数值为百分比。

表 3　不同算法在 MPII 数据集上的结果

模型	参数/M	*GFLOPs*	*PCKh*/%	*PCKh*@0.1/%
ShuffleNetV2[15]	7.6	1.7	82.8	20.5
Small HRNet[20]	1.3	0.7	80.2	—
MobileNetV2[27]	9.6	1.9	85.4	23.5
EfficientPoseI[36]	0.7	1.7	85.2	26.5
Lite-HRNet-18[21]	1.1	0.2	86.1	29.5
Lite-HRNet-30[21]	1.8	0.4	87.0	31.3
Dite-HRNet-18[23]	1.1	0.2	87.0	31.1
Dite-HRNet-30[23]	1.8	0.4	87.6	31.7
LH-HRNet-18 (ours)	1.6	0.4	87.4	31.9
LH-HRNet-30 (ours)	2.6	0.6	88.1	32.6

Small HRNet[20]，Lite-HRNet[21]，Dite-HRNet[23]，MobileNetV2[27]，EfficientPoseI[36] 相比，本文提出的 LH-HRNet-30 网络的 *PCKh* 和 *PCKh*@0.1 值分别为 88.1% 和 32.6%，是表 3 中性能最好的算法。

3）消融实验

为了进一步验证本文提出的算法性能，本文分别在 COCO val2017 和 MPII 数据集进行消融实验。其中 Proposed (A)表示 Dite-HRNet-18 作为基干网络，采用加强通道信息交

互的深度可分离卷积模块进行特征提取的算法;Proposed(B)表示在 Proposed(A)基础上,采用空洞卷积补全操作进行特征提取;Proposed(C)表示本文提出的网络结构。表 4 与 5 的实验结果表明 DCC 操作,PEM 模块和 IDSC 模块在不同程度上提高了算法性能,且同时使用三者时性能更好,说明每个改进模块对算法性能都有一定的提升。

表 4 算法在 COCO val2017 数据集上的消融实验

模型	参数/MB	*GFLOPs*	*AP*/%	AP^{50}/%	AP^{75}/%	AP^{M}/%	AP^{L}/%	*AR*/%
Dite-HRNet-18[23]	1.1	0.4	69.0	88.0	76.0	65.5	75.5	75.0
Proposed(A)	1.6	0.6	69.5	88.6	76.8	66.0	76.0	75.9
Proposed(B)	1.6	0.6	70.5	88.7	76.7	66.3	76.8	76.1
Proposed(C)	1.6	0.7	70.3	88.4	77.5	66.8	76.8	75.8

表 5 算法在 MPII 数据集上的消融实验 %

模型	*Head*	*Shoulder*	*Elbow*	*Wrist*	*Hip*	*Knee*	*Ankle*	*PCKh*	*PCKh*@0.1
Dite-HRNet-18[23]	96.2	94.3	87.0	80.7	87.4	81.9	77.7	87.0	31.1
Proposed(A)	96.3	94.8	87.1	80.9	87.2	82.5	77.5	87.2	31.8
Proposed(B)	96.2	94.7	87.3	80.9	87.7	82.1	78.1	87.3	31.8
Proposed(C)	96.5	94.7	87.2	81.3	87.3	82.5	78.0	87.4	31.9

4)实时性分析

为验证算法的推理效率,将本文算法与 Dite-HRNet-18 算法的推理效率进行对比分析。具体为:输入大小为 *FPS* 的单张 MPII 数据集图片,经过模型预测后的人体关键点检测时间,其模型推理效果如表 6 所示,其中 Proposed(A)、Proposed(B)、Proposed(C)和表 4 的含义相同。从表 6 可以看出:本文所提出的算法对单张图片的预测时间耗时稍微增加了一点,但是所提算法中的各个改进模块对提高检测精度方面都有积极的作用。

5)训练过程与可视化分析

图 9 表示模型整体训练过程中损失下降和准确率上升图。根据图中线条的整体趋势可以发现模型整体收敛情况良好,在 Warm-up 训练策略调整学习率后,模型收敛的效果更好。

表 6 模型实时性比较

模型	单张图片检测时间/ms	*FPS*	*GFLOPs*	参数/MB	*PCKh*/%	*PCKh*@0.1/%
Dite-HRNet-18[23]	7.5	133	0.2	1.1	87.0	31.1
Proposed(A)	8.0	124	0.3	1.6	87.2	31.8
Proposed(B)	8.0	124	0.4	1.6	87.3	31.8
Proposed(C)	8.1	122	0.4	1.6	87.4	31.9

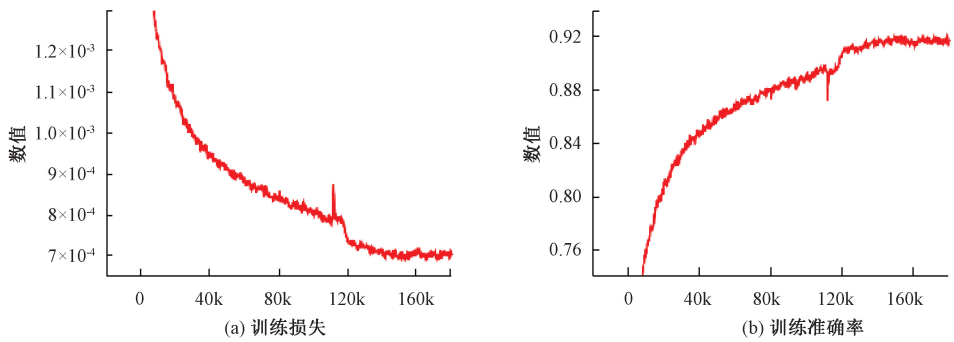
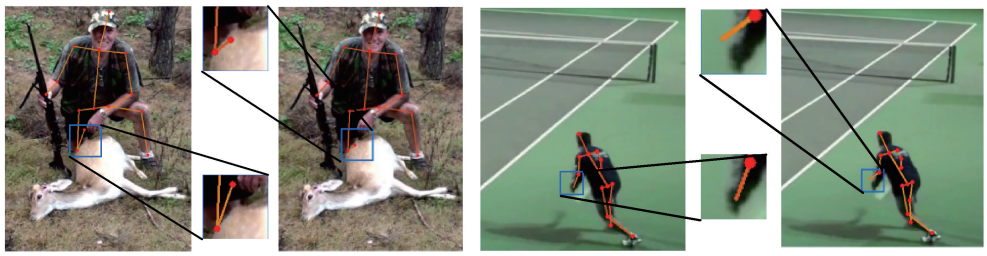


图 9 模型训练的收敛效果

模型检测到人体关键点热力图后,也可将其作为蒙版重新覆盖在原始图片上,并根据人体结构进行关键点之间的连接,实现人体姿态的可视化。如图 10(a)和(b)所示,每个子图中的左边部分表示 LH-HRNet 模型检测效果,子图中的右边部分表示 Dite-HRNet 模型检测效果,从中可以看到,本文算法 LH-HRNet 检测到的关键点位置定位更准确。



(a) 日常生活图　　(b) 体育运动图

图 10　人体姿态估计算法 LH-HRNet 和 Dite-HRNet 可视化结果

4　结　　论

为了解决轻量级卷积神经网络对高分辨特征信息挖掘不充分的问题，本文提出了一种基于特征增强的高分辨率人体姿态估计网络提取特征信息。由于池化增强和通道信息增强模块的有效性和空洞卷积核补全操作既能解决空洞卷积缺陷又能增加卷积操作的感受野，因此本文提出的网络在 COCO2017 和 MPII 人体姿态估计数据集上都取得较好的检测效果，本文提出的 LH-HRNet-30 网络比 Dite-HRNet-30 网络有比较显著的性能提升，特别是输入图片大小为 384×288 比 256×192 的精度提高更为明显。以上结果表明本文算法对较大图片的提取特征能力更强，即对高分辨率图片的提取能力更强。此外，本文提出的 LH-HRNet-18 实现了接近于 Dite-HRNet-30 算法的性能，但是参数量普遍小于 Dite-HRNet-30 算法。未来工作将所提出的轻量级模块进一步扩展到其它领域的高分辨图片网络中，从而有效的提取浅层特征。

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.

[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.

[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.

[4]　尹智帅，钟恕，聂琳真，等. 基于人体姿态估计的分心驾驶行为检测[J]. 中国公路学报，2022，35(6)：312-323.

[5]　张小娜，吴庆涛. 基于深度学习的自顶向下人体姿态估计算法[J]. 电子测量技术，2021，44(9)：105-109.

[6]　ALEXANDER T，CHRISTIAN S. DeepPose：Human pose estimation via deep neural networks[C]. 2014 IEEE Conference on Computer Vision and Pattern Recognition，USA：IEEE Computer Society，2014：1653-1660.

[7]　TOMPSON J 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 system，2014，1：1799-1807.

[8]　HUANG J J，ZHU Z，GUO F，et al. The devil is in the details：Delving into unbiased data processing for human pose estimation[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition，IEEE Computer Society，2020：5700-5709.

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

[10]　SUN K，XIAO B，LIU D，et al. Deep high-resolution representation learning for human pose estimation[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition，IEEE Computer Society，2019：5693-5703.

[11]　CHEN Y，WANG Z，PENG Y，et al. Cascaded pyramid network for multi-person pose estimation[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition，IEEE Computer Society，2018：7103-7112.

[12]　ZHANG L X，HUANG W T，WANG C L，et al. Improved multi-person 2D human pose estimation using attention mechanisms and hard example mining[J]. Sustainability，2023，15(18)：13363-13380.

[13]　DONG X Q，WANG X C，LI B J，et al. YH-Pose：Human pose estimation in complex coal mine scenarios[J]. Engineering Applications of Artificial Intelligence，2024，127：107338-107351.

[14]　HU J，SHEN L，ALBANIE S，et al. Squeeze-and-excitation networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligenc，2020，42(8)：2011-2023.

[15]　MA N，ZHANG X，ZHENG H T，et al. Shufflenet v2：Practical guidelines for efficient CNN architecture design[C]. Proceedings of the European Conference on

Computer Vision. September 8-14, 2018, Munich, Germany: Springer, 2018: 116-131.
[16] 万乐玲,行鸿彦. 基于微信小程序的人体尺寸测量系统[J]. 电子测量技术, 2021, 44(22): 141-147.
[17] 林之钊,黄惠. 从单幅图像生成人物三维动画[J]. 计算机辅助设计与图形学学报, 2022, 34(9): 1341-1350.
[18] 付心仪,蔡天阳,薛程,等. 基于 BGRU-FUS-NN 神经网络的姿态情感计算方法研究[J]. 计算机辅助设计与图形学学报, 2020, 32(7): 1070-1079.
[19] 丁静,舒祥波,黄捧,等. 基于多模态多粒度图卷积网络的老年人日常行为识别[J]. 软件学报, 2023, 34(5): 2350-2364.
[20] 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.
[21] YU C, XIAO B, GAO C, et al. Lite-hrnet: A lightweight high-resolution network[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE Computer Society Press, 2021.
[22] CUI C, GAO T, WEI S, et al. PP-LCNet: A lightweight CPU convolutional neural network [J/OL]. (2021-09-17)[2023-09-01]. https://arxiv.org/abs/2109.15099.html.
[23] LI Q, ZHANG Z, XIAO F, et al. Dite-hrnet: Dynamic lightweight high-resolution network for human pose estimation [C]. Proceedings of the International Joint Conference on Artificial Intelligence. November 6-9, 2022, Shenzhen, China: Morgan Kaufmann, 2022: 1095-1101.
[24] 朱翠涛,李博. 基于高分辨率网络的人体姿态估计[J]. 中南民族大学学报(自然科学版), 2023, 42(2): 229-237.
[25] 李丽,张荣芬,刘宇红,等. 基于多尺度注意力机制的高分辨率网络人体姿态估计[J]. 计算机应用研究, 2022, 39(11): 3487-3491, 3497.
[26] ZHANG X, ZHOU X, LIN M, et al. Shufflenet: An extremely efficient convolutional neural network for mobile devices [C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, 2018: 6848-6856.
[27] SANDLER M, HOWARD A, ZHU M, et al. Mobilenetv2: Inverted residuals and linear bottlenecks[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, 2018: 4510-4520.
[28] ZHOU D, HOU Q, CHEN Y, et al. Rethinking bottleneck structure for efficient mobile network design[C]. Proceedings of the European Conference on Computer Vision. August, 2020.
[29] YU F, KOLTUN V. Multi-Scale Context Aggregation by Dilated Convolutions[J/OL](2015-11-23)[2023-09-01]. https://arxiv.org/abs/1511.07122.html.
[30] CHEN L C, PAPANDREOU G, KOKKINOS I, et al. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs[J]. Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(4): 834-848.
[31] WANG P, CHEN P, YUAN Y, et al. Understanding convolution for semantic segmentation [C]. Proceedings of 2018 IEEE Winter Conference on Applications of Computer Vision, IEEE Computer Society, 2018: 1451-1460.
[32] DING X, GUO Y, DING G, et al. Acnet: strengthening the kernel skeletons for powerful cnn via asymmetric convolution blocks[C]. Proceedings of the International Conference on Computer Vision, IEEE Computer Society, 2019: 1911-1920.
[33] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft coco: common objects in context [C]. Proceedings of the European Conference on Computer Vision, Springer International Publishing, 2014: 740-755.
[34] 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/CVF Conference on computer Vision and Pattern Recognition, IEEE Computer Society, 2014: 3686-3693.
[35] XIAO B, WU H, WEI Y. Simple baselines for human pose estimation and tracking[C]. Proceedings of the European Conference on Computer Vision, Springer, 2018: 466-481.
[36] GROOS D, RAMAMPIARO H, IHLEN E A F. Efficientpose: Scalable single-person pose estimation[J]. Applied Intelligence, 2021, 51(4): 2518-2533.

作者简介

谢唯嘉,硕士研究生,主要从事计算机视觉方面的研究。
E-mail:6120210299 @mail.jxust.edu.cn

易见兵,博士,副教授,硕士生导师,主要从事计算机视觉、高性能计算等方面的研究。
E-mail: yijianbing8@jxust.edu.cn

曹锋,博士,讲师,硕士生导师,主要从事智能信息处理、自动定理证明、程序可信性验证等方面的研究。
E-mail: caofeng19840301@163.com

李俊,博士,讲师,硕士生导师,主要从事计算机视觉、自然语言处理、机器学习方面的研究。
E-mail: lijun17@jxust.edu.cn