E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP391; TP183　　**文献标识码**: A　　**文章编号**: 1006-8961(2024)05-1408-13

论文引用格式:Gao K, Li W G, Shu Y, Ge Y K and Wang Z G. 2024. Lightweight high-resolution human pose estimation combined with densely connected network. Journal of Image and Graphics, 29(05):1408-1420(高坤, 李汪根, 束阳, 葛英奎, 王志格. 2024. 结合密集连接的轻量级高分辨率人体姿态估计. 中国图象图形学报, 29(05):1408-1420)[DOI:10.11834/jig.230228]

结合密集连接的轻量级高分辨率人体姿态估计

高坤,李汪根*,束阳,葛英奎,王志格

安徽师范大学计算机与信息学院,芜湖　241002

摘　要:**目的**　为了更好地实现轻量化的人体姿态估计,在轻量级模型极为有限的资源下实现更高的检测性能。基于高分辨率网络(high resolution network,HRNet)提出了结合密集连接网络的轻量级高分辨率人体姿态估计网络(lightweight high-resolution human estimation combined with densely connected network,LDHNet)。**方法**　通过重新设计HRNet中的阶段分支结构以及提出新的轻量级特征提取模块,构建了轻量高效的特征提取单元,同时对多分支之间特征融合部分进行了轻量化改进,进一步降低模型的复杂度,最终大幅降低了模型的参数量与计算量,实现了轻量化的设计目标,并且保证了模型的性能。**结果**　实验表明,在MPII(Max Planck Institute for Informatics)测试集上相比于自顶向下的轻量级人体姿态估计模型LiteHRNet,LDHNet仅通过增加少量参数量与计算量,平均预测准确度即提升了1.5%,与LiteHRNet的改进型DiteHRNet相比也提升了0.9%,在COCO(common objects in context)验证集上的结果表明,与LiteHRNet相比,LDHNet的平均检测准确度提升了3.4%,与DiteHRNet相比也提升了2.3%,与融合Transformer的HRFormer相比,LDHNet在参数量和计算量都更低的条件下有近似的检测性能,在面对实际场景时LDHNet也有着稳定的表现,在同样的环境下LDHNet的推理速度要高于基线HRNet以及LiteHRNet等。**结论**　该模型有效实现了轻量化并保证了预测性能。

关键词:人体姿态估计;轻量级网络;密集连接网络;高分辨率网络;多分支结构

Lightweight high-resolution human pose estimation combined with densely connected network

Gao Kun, Li Wanggen*, Shu Yang, Ge Yingkui, Wang Zhige

School of Computer and Information, Anhui Normal University, Wuhu　241002, China

Abstract: **Objective** Human pose estimation is a technology that can be widely used in life. In recent years, many excellent high-precision methods have been proposed, but they are often accompanied by a very large model scale, which will encounter the problem of computing power bottleneck in application. Whether for model training or deployment, large models require considerable computing power as the basis. Most of them have low computing power. Similarly, for the scenes in daily life, the equipment needs further applicability and detection speed of the model, which is difficult to achieve by large models. Given such requirements, lightweight human pose estimation has become a hot research field. The main problem is how to achieve high detection accuracy and fast detection speed under the extremely limited number of resources. Lightweight models will inevitably fall into a disadvantage in detection accuracy compared with large models.

收稿日期:2023-05-31;**修回日期**:2023-07-31;**预印本日期**:2023-08-07

*通信作者:李汪根　1346858911@qq.com

基金项目:国家自然科学基金项目(61976006)

Supported by:National Natural Science Foundation of China(61976006)

However, fortunately, from many studies in recent years, the lightweight model can also achieve higher detection accuracy than large ones. A good balance can be reached between them. **Method** Based on a high-resolution network (HRNet), a lightweight high-resolution human pose estimation network combined with a dense connection network (LDHNet) was proposed. First, dense connection and multi-scale were integrated to construct a lightweight and efficient feature extraction unit by redesigning the stage branch structure in HRNet and proposing a new lightweight feature extraction module. Then, the feature extraction module is composed of modules similar to the pyramid structure, and the dilated convolution of three scales is used to obtain a wide range of feature information in the feature map by stacking the multi-layer feature extraction modules and fusing the output of each layer. The concatenation of the output feature map of the feature extraction module is to reuse the feature map and fully extract the information contained in the feature map. These two points can make up for the problem of insufficient utilization of feature information that may exist in the lightweight model and use limited resources to achieve high feature extraction performance. Second, a wide range of cross-branch feature information interactions exists in the original HRNet structure, including feature fusion and the generation of new branches in each stage. LDHNet replaces the convolution in this process with the depthwise separable convolution by changing the size of the feature map through convolution downsampling or upsampling to add with other branch feature maps for feature fusion. This case further reduces the number of parameters of the model based on almost no loss of detection performance. In addition, LDHNet improves the original data preprocessing module and uses the double-branch form to fully extract the information from the original image. Experiments show that considerable information from the original image is of great help to improve the detection performance of the model. LDHNet also uses coordinate attention to reinforce spatial location information. **Result** After the above improvements, the size of the model has been greatly reduced to less than one-tenth of the original HRNet. Although some gap still exists between the model size and the current smallest lightweight model LiteHRNet, the design of the lightweight model is not only concerned with the size of the model. This study mainly compares LDHNet with mainstream lightweight models. Through experimental verification on two mainstream datasets of human pose estimation MPII and common objects in context(COCO) dataset and comparison with the current mainstream methods, the following conclusions can be obtained. Compared with the top-down lightweight human pose estimation LiteHRNet on the MPII test set, the average prediction accuracy of LDHNet is improved by 1.5% by only adding a small number of parameters and calculations. The results on the COCO validation set show that compared with LiteHRNet, the average detection accuracy of LDHNet is improved by 3.4%. Compared with the improved DiteHRNet of LiteHRNet, the detection accuracy of LDHNet is improved by 2.3%. Compared with the HRFormer fused with Transformer, the detection accuracy of LDHNet is the same when the scale is smaller. The experimental results on public datasets show that LDHNet achieves excellent results in model lightweight. LDHNet achieves a very good balance between lightweight and model detection performance. For lightweight human pose estimation, the performance of LDHNet is similar to that of Transformer. In addition to experimental verification in the public data set, this study also tests the inference speed and detection accuracy of the model in the actual scene. Compared with the original HRNet, LDHNet has a significant improvement in the detection speed under GPU acceleration. Compared with other lightweight methods, such as LiteHRNet, LDHNet can make full use of hardware resources. The results of the actual test show that LDHNet has a stable performance in the face of the actual scene. The above experimental results show that LDHNet has achieved the design expectations. Moreover, using extremely limited resources, compared with other lightweight human pose estimation models, LDHNet can make full use of all the computing power regardless of the level of hardware computing power. The detection accuracy is greatly improved, and the inference speed of the model is also significantly improved. **Conclusion** For LDHNet, some problems also need to be solved, mainly in that the reasoning speed of the model has not reached the level matching the improvement of the model in terms of lightweight. Follow-up works can focus on how to improve the reasoning speed of the model, make the training and reasoning of the model glass, and use the method of emphasizing parameters for reference to further improve the model. Thus, LDHNet can be further competent for the needs of actual production and life.

Key words: human pose estimation; lightweight network; densely connected network; high resolution network; multi-branch structure

0 引 言

2D人体姿态估计是计算机视觉领域几个重要的研究方向之一，其主要目标是从场景中提取人体的关节点坐标，广泛应用在各类场景中（Luvizon等，2018），在包括动作识别、人机交互以及虚拟现实等应用场景中人体姿态估计都扮演着举足轻重的角色（孔英会 等，2023）。人体姿态估计主要分为两大类方法：一类是自顶向下的方法（top down），这类方法首先从场景中检测出单个人体目标，再对所检测出的人体目标区域进行姿态估计；另一类是自底向上的方法（bottom up），这类方法直接从场景中检测所有的关节点，再将所检测出的关节点按一定规则组装出人体结构。得益于将多人姿态估计转换为单人姿态估计，在性能上自顶向下的方法表现更好并且具有更好的鲁棒性（邓益侬 等，2019）。

堆叠沙漏网络（stacked hourglass networks，SHN）（Newell等，2016）通过堆叠多个沙漏结构的单元对特征信息进行反复提取，由于对特征图的多次下采样和上采样，存在特征信息丢失的问题；级联金字塔网络（cascaded pyramid network，CPN）（Chen等，2018）引入特征金字塔对特征进行提取并使用修正网络对特征金字塔的输出进行修正，兼顾了多尺度的特征信息，在一定程度上弥补了由于特征信息丢失所造成的性能损失；简单基线网络（simple base line，SBL）（Xiao等，2018）仅通过最简单的下采样和转置卷积上采样就取得了优异的性能表现；高分辨率网络（high resolution network，HRNet）（Sun 等，2019）通过过往的研究认识到高分辨率特征图对于姿态估计的重要性，所以选择构建多个网络分支，将高分辨率分支一直保留下来，构建出一个强劲的网络结构，不仅在姿态估计领域取得了优异的表现，在包括实例分割、分类在内的多个问题中都取得了相对优异的性能表现。ViTPose（vision transformer pose estimation）（Xu等，2022）通过引入Vision Transformer解决人体姿态估计问题，使用简单的单分支模型结构就取得了远超HRNet的性能表现，验证了最原始的Vision Transformer就具有相比于卷积神经网络（convolutional neural network，CNN）在视觉任务中更强劲的学习能力，不需要复杂的模型设计就可以提取到足够高质量的特征信息。

在人体姿态估计研究日渐成熟的过程中，伴随着模型性能的不断提升，模型的规模也不断增大，动辄上千万的参数量无论对于模型的训练还是部署应用都需要相对高昂的成本，与实际生活中常见的低算力设备的高适用性和轻量快速需求相悖，所以轻量化的人体姿态估计成为一个重要的研究方向（蔡哲栋 等，2021），近些年有基于简单基线网络（SBL）提出的轻量级简单人体姿态估计网络（simple and lightweight human pose estimation，LPN）（Zhang 等，2020），通过简化SBL的结构以及使用深度可分离卷积代替原本瓶颈结构缩小了模型的规模，并且在推理速度方面与其他方法相比具有较大优势；基于高分辨率网络（HRNet）提出的Lite-HRNet（Yu 等，2021），极大地降低了模型的参数量和计算量；高分辨率Transformer（HRFormer）（Yuan等，2021）构建了一个高分辨率的Transformer网络，在模型足够轻量化的同时，性能表现同样优异。

经过对人体姿态估计研究领域的深入调研，本文选择设计一个轻量级的人体姿态估计网络即结合密集连接的轻量级高分辨率网络（lightweight high-resolution human estimation combined with densely connected network，LDHNet），在设计过程中，面对的主要问题是如何利用极为有限的资源达到更高的检测性能，本文将HRNet作为基线，通过重新设计网络中的各阶段结构，融入密集连接和多尺度特征提取，充分利用网络各阶段中的特征图，避免对其所包含信息的浪费，同时对模型中占用资源较大的特征融合部分进行简化，使其变得足够轻量化，在完成上述这些设计工作后通过实验与其他方法进行对比，验证LDHNet达到了轻量化模型设计预期。

与此同时，还有如下一些背景工作，这里进行简单介绍：

1）密集连接网络。密集连接网络通过前向传播的方式将每层都与后续层相连接，使得每层特征图都包含其前置所有层的信息，不仅能够最大程度对特征图进行复用从而加强特征信息的传播，也有助于训练过程中梯度的反向传播，DenseNet（Huang等，2017）通过将所有前面层与当前层相拼接的方式构建网络，这样的连接方式虽然提升了网络性能，但会使网络的复杂度急剧增加，在DenseNet的基础上，VoVNet（Lee等，2019）提出仅保留DenseNet中所有前置层与最后一层的密集连接，在保持密集连接

网络优势的同时几乎不增加网络的复杂度。对于轻量级网络而言，能够提高对特征信息的利用效率，是一个提升性能的有效途径。

2）高分辨率网络。人体姿态估计是对位置信息极为敏感的任务，经过网络的反复下采样得到低分辨率特征图，会损害空间维度的信息，从而导致难以预测到准确的位置信息，为了弥补空间信息上的缺失，高分辨率网络构建了一个多分支网络，始终保持高分辨特征图分支，并逐步引入下采样增加低分辨率特征图分支与高分辨率特征图分支并行连接，同时在多个分辨率特征图分支之间进行反复的信息交互，以提升特征图的表征能力，最终将最高分辨率特征图分支的特征图作为输出。

高分辨率网络结构提供了更好的特征提取能力，使得轻量级网络可以更高效地从输入数据中提取有用的信息。具体来说，高分辨率网络结构通过增加网络的深度和宽度，使得网络可以学习到更多的特征（Sun 等，2019），对于视觉类任务而言这能够极大地提升模型的性能。

3）轻量级网络。轻量级 CNN 网络是一种专门针对移动设备和嵌入式设备等资源有限的设备而设计的卷积神经网络。相比于传统的 CNN 网络，轻量级 CNN 网络通常具有更少的参数数量、更小的模型体积和更低的计算复杂度，同时仍能保持较高的准确率和优秀的性能表现。这使得轻量级 CNN 网络成为许多应用场景中的理想选择，例如移动端的人脸识别、图像分类和物体检测等。

常见的轻量级 CNN 网络如 MobileNet（Howard 等，2017）、ShuffleNet（Ma 等，2018）、SqueezeNet（Iandola 等，2016）、ESPNet（efficient spatial pyramid of dilated convolutions for semantic segmentation）（Mehta 等，2019）等，它们都采用了一系列的优化策略来减少参数数量和计算复杂度。MobileNet 通过使用深度可分离卷积代替传统的卷积操作来减少参数数量；ShuffleNet 则采用了组卷积和通道混洗等技术来降低计算复杂度。这些优化策略在保证模型性能的前提下，有效地提高了轻量级 CNN 网络在移动端的实时性和运行效率。将轻量级网络与高分辨率网络结构相结合，能够兼顾模型的轻量化和高分辨率网络结构强劲的特征提取性能。

1 本文方法

1.1 LDHNet

本文提出的 LDHNet 以高分辨率网络（HRNet）作为基线，在完成对输入特征图的预处理后主要进行 4 个阶段，第 1 阶段仅存在一个最高分辨率分支，在后续的 3 个阶段中在每个阶段的初始会由当前阶段内最低分辨率分支进行下采样生成一个新分支，在多分支的阶段内会反复进行不同分支之间的特征信息交互，在第 4 阶段结束后仅输出最高分辨率分支的特征图用于预测关键点，其整体结构如图 1 所示，具体组成如表 1 所示。由表 1 可以看出，LDHNet 在第 1 阶段中仅采用 GMCneck 结构，通过简单的串联结构实现，在后续阶段中使用 GWblock 结构与跨

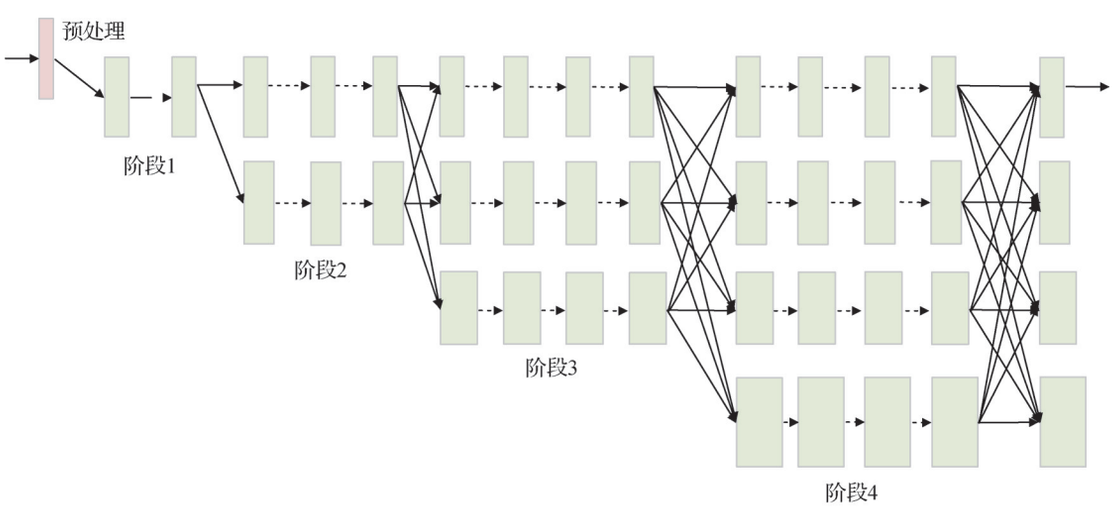

图 1　LDHNet 结构图

Fig. 1　Overall structure diagram of LDHNet

分支特征交互部分实现，GWblock结构使用密集连接的方式将多组GMConv相互联系起来，跨分支特征交互部分对特征图进行变换并与其他分支特征图相加进行特征交互。另外在每个阶段中生成新分支时，会由现有的多个分支的特征图进行下采样融合生成，生成的新分支特征图尺寸为最小分辨率分支特征图尺寸的1/4，通道数相较最小分辨率的分支增加一倍。

表1　LDHNet组成

Table 1　The composition of LDHNet

阶段	操作	堆叠数目	分支数目	通道数目	是否多尺度输出
1	GMCneck	1	1	64	是
2	GWblock+Fuse	1	2	32，64	是
3	GWblock+Fuse	4	3	32，64，128	是
4	GWblock+Fuse	3	4	32，64，128，256	否

1.2　Stem模块

HRNet中原本的Stem模块仅通过串联两层3×3的卷积对特征图进行变换，在一定程度上损失了特征信息，受PeleeNet（Wang等，2018）提出的双分支特征提取模块的启发，本文设计了一个结合密集连接的Stem模块，从而保留更多来自原始特征图中的信息，Stem模块的结构如图2所示。

将原HRNet中的Stem的第2个3×3卷积替换为卷积与池化两个分支对特征图进行下采样，并在之后将两个输出的特征图进行通道维度的拼接，随后使用1×1的卷积对特征图进行维度变换还原尺寸，通过实验对比，相比原HRNet的Stem模块，LDHNet中的Stem模块在不增加参数量和计算量的前提下，提高了模型的检测性能。

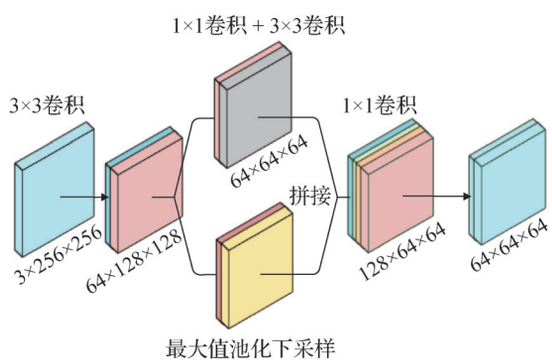

图2　Stem模块

Fig. 2　Stem module

1.3　GMCneck

GMCneck由3组GSConv模块进行串联组成，GSConv的结构如图3所示，首先对特征图进行通道维度的压缩来有效降低所需的参数量和计算量，本文在GSConv模块中设置的压缩倍率为2，即将特征图通道数压缩为原本的1/2，随后使用深度卷积（Sandler等，2018）与ECA（efficient channel attention）注意力机制（Wang等，2020）进行特征提取，最后对特征图进行通道维度的拼接并还原为所需的特征图尺寸。本文在LDHNet中使用3组GSConv模块进行串联构成GMCneck结构。

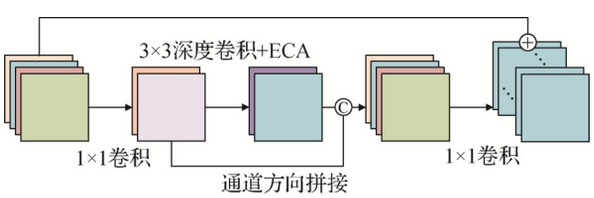

图3　GSConv模块

Fig. 3　GSConv module

1.4　GMConv模块

通过对GSConv模块的思考，轻量级特征提取模块在特征提取过程中会难以兼顾更大尺度范围中的信息，从而导致性能表现不佳，所以本文在GSConv模块的基础上结合ESPNet的思想（Mehta等，2019），提出了GMConv模块，同时兼顾轻量级与更大尺度范围内的特征信息，GMConv模块的结构如图4所示。

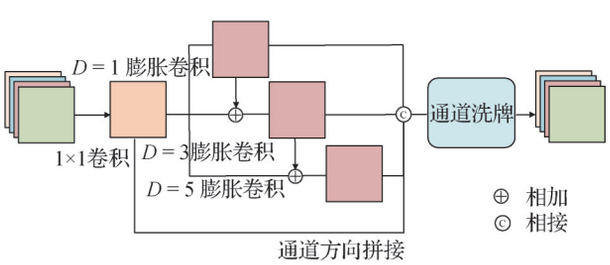

图4　GMConv模块

Fig. 4　GMConv module

首先，仍然采用GSConv对特征图进行通道压缩的方式来降低所需的参数量和计算量，在这里本文选择的压缩倍率为4，即通道数目压缩至原特征图的1/4，具体为

$$\boldsymbol{X}_T = \varphi(\boldsymbol{X}) \tag{1}$$

式中，$\boldsymbol{X}_T$代表对输入特征图进行通道维度压缩后的特征图，$\varphi(\boldsymbol{X})$代表对特征图$\boldsymbol{X}$进行通道维度的压缩。

随后，使用3组膨胀率不同的膨胀卷积进行特征提取，并将膨胀率更小的卷积输出的特征图与下一组膨胀卷积的输出特征图相加，以此类推得到3组输出特征图，具体为

$$\boldsymbol{X}_{T(i)} = D_i(\boldsymbol{X}_T) + \boldsymbol{X}_{T(i-1)},\ \ i = 1,2,3 \tag{2}$$

式中，$\boldsymbol{X}_{T(i)}$代表第i个分支对特征图进行处理的输出特征图，需要注意的是默认$\boldsymbol{X}_{T(o)}$为空，$D_i(\boldsymbol{X}_T)$代表对特征图$\boldsymbol{X}_T$进行第i组膨胀卷积，膨胀率依次为1，3，5。

最后，将4组特征图进行通道维度的拼接并使用通道洗牌，得益于GMConv首先将特征图通道数目压缩至原本的1/4，在对特征图进行重新拼接时刚好能够还原回原尺寸，无需再使用1×1卷积进行通道维度的调整，减少了不必要的操作，有效降低了参数量和计算量，具体为

$$\boldsymbol{X}_{\text{out}} = Concat[\boldsymbol{X}_T, \boldsymbol{X}_{T(1)}, \boldsymbol{X}_{T(2)}, \boldsymbol{X}_{T(3)}] \tag{3}$$

式中，$\boldsymbol{X}_{\text{out}}$代表输出特征图，$Concat$代表将特征图沿通道维度进行拼接并使用通道洗牌来打乱特征图的通道维度顺序，通过实验对比，使用GMConv模块进行特征提取，相比于GSConv模块，有着更加轻量和检测精度更高的优势。

1.5 GWblock

基于密集连接网络的思想，本文设计了GWblock结构，由两组GMConv基本模块作为密集单元以及最后特征图拼接后的1×1卷积4个部分构成，通过堆叠并且密集连接的方式相互联系，GWblock结构的示意图如图5所示。

GWblock使用密集连接将输入特征图以及中间每层的输出与最后一个GMConv的输出相拼接，通过密集连接的方式，GWblock的最后一层会包含其所有前置层中的特征信息，使得中间的所有特征图中的信息都能得到充分利用，并且在训练过程中梯度可以快速地反向传播，提高网络的训练效率，另外由于仅保留了所有中间层与最后一层之间的密集连接，所以仅从增加的连接角度来看，参数量和计算量并没有相比直接对GMConv模块串联明显提升，在提升模型性能的同时也保证了模型的轻量化。

首先，在GWblock中的第1个GMConv模块在负责特征提取的同时还负责进行通道数目变换，将输入特征图的通道数转换为GWblock结构最终输出的通道数，后续的两个GMConv不再对特征图的通道数目进行调整，具体为

$$\boldsymbol{Y}_1 = \varphi(\boldsymbol{x}_{\text{in}}) \tag{4}$$

$$\boldsymbol{Y}_i = \varphi(\boldsymbol{Y}_{i-1}) \tag{5}$$

式中，$\boldsymbol{x}_{\text{in}}$代表输入GWblock的特征图，$\boldsymbol{Y}_i$代表第$i$个GMConv的输出。在完成3组GMConv模块后要将所有中间层特征图包括$\boldsymbol{x}_{\text{in}}$进行拼接，最后对拼接的特征图进行1×1卷积，将通道数目转换至与特征图$\boldsymbol{x}_{\text{in}}$通道数相同，具体为

$$\boldsymbol{x}_{\text{out}} = \varphi(Concat[\boldsymbol{x}, \boldsymbol{Y}_1, \boldsymbol{Y}_2, \boldsymbol{Y}_3]) \tag{6}$$

式中，$\boldsymbol{x}_{\text{out}}$代表输出的特征图，$Concat$代表沿通道方向拼接特征图，$\varphi$代表使用的线性变换。通过实验的验证使用GWblock替换HRNet中的基础单元，模型的参数量和计算量都得到了大幅降低，并且模型的检测精度也得到保证。

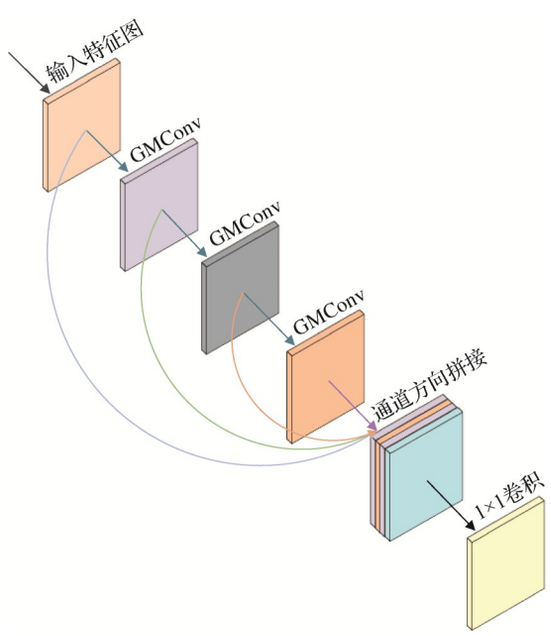

图5 GWblock结构

Fig. 5 GWblock structure

1.6 特征融合

HRNet中多个分辨率分支中的特征进行多次融合，每个分支都会接受来自其他所有分辨率分支中

的信息，用来增强该分支内特征图的表征能力，在HRNet中对由高分辨率分支向低分辨率分支传递信息时采用的是步长为2的3×3卷积来实现，若两个分支之间的跨度超过1，那么执行多次下采样，相同分辨率分支时不采取操作直接映射，由低分辨率向高分辨率分支传递信息时采用的系数为2的邻临近上采样，若两个分支之间的跨度超过1同样进行多次上采样，最后将多个分辨率传递的信息相加得到该分支的新特征图。

本文对HRNet中的跨分支交互过程进行了一些改进，其中包括在每个阶段内的跨分支特征交互以及生成新分支时的特征交互两个部分，如图6所示，可以看到将HRnet中特征融合部分的由高分辨率向低分辨率分支传递信息的下采样部分使用的普通的3×3卷积替换为3×3的深度可分离卷积，并在其中加入了坐标注意力机制（Hou等，2021），对于视觉类任务，对空间信息的敏感度相对较高，加入坐标注意力机制有助于将更重要的空间位置信息着重提取出来，经过以上改动，大幅削减了模型的规模，对模型的轻量化帮助显著。

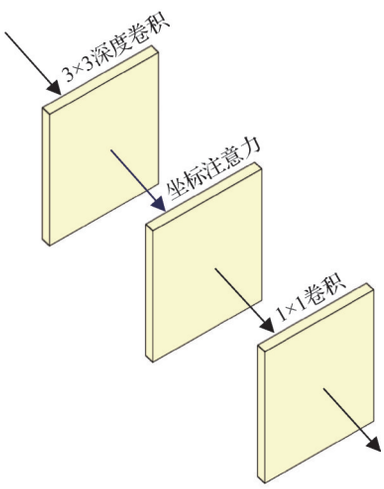

图6　特征融合

Fig. 6　Feature fusion

2　实验分析

2.1　数据集介绍

本文选择了两个数据集进行模型改进效果的验证（高坤 等，2022），分别为COCO2017（common object in context）数据集和MPII（Max Planck Institute for Informatics）数据集。COCO2017数据集（Lin等，2014）是微软提供的可用于多种计算机视觉任务的数据集，其涵盖方向包括目标检测、实例分割和姿态估计等，数据集中包含约330 K幅图像，其中超过200 K幅图像具有标注信息，COCO2017数据集划分为训练集、验证集和测试集，对于人体姿态估计任务主要使用的是COCO2017数据集中包含人体关键点标注的图像，总计约250 K个人体目标，可使用训练集中的约57 K幅图像进行训练，使用验证集中的约5 K幅图像中进行验证，最后使用测试集中的20 K幅图像进行测试。 COCO2017数据集在人体姿态估计领域所使用的评价指标是关键点相似性（object keypoint similarity，OKS），具体为

$$OKS=\frac{\sum_i\left[\exp\left(-\frac{d_{pi}^2}{2s_p^2\sigma_i^2}\right)\delta(v_i>0)\right]}{\sum_i\left[\delta(v_i>0)\right]} \tag{7}$$

式中，d_{pi}表示模型所检测的关键点与数据集中所标注的关键点之间的欧氏距离，v_i表示关键点在数据集中的标注信息，s_p表示尺度因子，σ_i表示归一化因子，$\delta(v_i>0)$表示当$v_i>0$时为1，否则为0，从而只对可见关键点进行计算。OKS的取值在0与1之间，值越大代表检测的准确度越高，实验中的评估标准是按照OKS取值确定的，评估标准包括AP（OKS=0.5，…，0.95时10个位置的预测准确度均值）、AP^{50}（OKS=0.50时的预测准确度）、AP^{75}（OKS=0.75时的预测准确度）、AP^{M}（中型物体预测准确度）、AP^{L}（大型物体预测准确度）以及AR（在OKS=0.50，…，0.95时10个位置的平均召回准确度）6项指标。

MPII数据集（Andriluka等，2014）是一个主要针对人体姿态估计任务的数据集，包含超过40 K人体目标的约25 K幅图像，并且这些人体目标均标注了16个人体关键点，它的主要评估指标是PCKh（percentage of correct keypoints），指预测正确的关键点的比例，将预测得到的关键点与标注关键点之间的归一化距离计算出来，若小于设定阈值α则认为预测正确，本文选择的评估标准为PCKh@0.5，即设定的阈值α=0.5，在评估过程中使用多个关键点进行评估，评估的关键点分别为头部（head）、肩膀（shoulder）、手肘（elbow）、腕部（wrist）、髋部（hip）、膝盖（knee）以及脚踝（ankle）。

2.2 实验环境设置

本文使用的实验平台为Ubuntu 20.04 LTS，CPU为i5-10400F，GPU为RTX 3060，显存大小为12 GB，深度学习框架为Pytorch1.10.1，软件平台为Python3.9，训练使用Adam优化器作为模型的优化器，设定的训练轮数与HRNet基线相同，为210轮，初始学习率设置为0.001，并且在第170轮和第200轮时将学习率分别缩减10倍。

由于实验数据集中的样本图像尺寸并不统一，首先需要对图像进行处理再进行训练，将图像中人体目标髋部作为中心，对图像进行剪裁，COCO数据集分别剪裁出256×192像素和384×288像素的图像，MPII数据集剪裁为256×256像素的图像，便于与其他方法进行对比，为方便对比实验，数据增强预处理方面采用与HRNet相同的策略，也与诸如LiteHRNet、DiteHRNet等基于HRNet改进的模型一致，在±45°范围内进行旋转以及在±35%范围内进行缩放。

2.3 实验验证

本文首先在COCO2017数据集上进行了实验验证，将LDHNet与其他经典主流方法以及最新轻量级模型进行参数量、计算量以及模型性能等方面的对比，在COCO验证集上的结果如表2所示，其中PT指标代表模型是否进行预训练。对于输入尺寸为256×192像素时，LDHNet的平均检测准确度达到了70.6%，在与同样作为轻量化模型的LPN的对比中，在参数量保持较大优势的同时模型检测准确度提升了0.2%，同时在与轻量级姿态估计方法LiteHRNet的对比中，虽然参数量和计算量有所提高，但是模型的平均检测准确度显著提升了3.5%，与对HRNet进行简单缩放轻量化的Small HRNet（Wang等，2021）进行对比，仅提升了一些模型规模，在检测精度方面就获得了巨大的提升，在与LiteHRNet的改进型DiteHRNet（Li等，2022）的对比中，模型的平均检测准确度同样有着2.4%的提升，并且在参数量和计算量更低一些的情况下与使用Transformer的HRFormer相比精度相近，在AP^{50}、AP^{M}指标上有较大优势；另外在与大型网络诸如SHN、CPN、SBL-50以及HigherHRNet（Cheng等，2020）等的对比中，参数量和计算量表现出极大优势，在与基线HRNet的对比中虽然平均检测准确度有所折损，但是AP^{50}相比HRNet有所提升，另外得益于更轻量化的结构，在对模型进行实时性测试时LDHNet相比HRNet有较大优势。

本文还在COCO test-dev数据集上对LDHNet进行了测试，结果如表3所示，相比于轻量级姿态估计方法LiteHRNet，在平均检测准确度方面与LiteHRNet相比可提高0.8%，COCO test-dev的测试表现

表2 COCO验证集实验结果对比

Table 2 Comparison of experimental results of COCO val set

方法	输入尺寸/像素	PT	Backbone	参数量/M	GFlops	AP/%	AP^{50}/%	AP^{75}/%	AP^{M}/%	AP^{L}/%	AR/%
SHN	256×256	N	SHN	25.1	14.3	66.9	–	–	–	–	–
CPN	256×192	Y	ResNet-50	27.0	6.2	68.6	–	–	–	–	–
SBL-50	256×192	Y	ResNet-50	34.0	8.9	70.4	88.6	78.3	67.1	77.2	76.3
HigherHRNet	256×192	Y	HRNet-W32	28.6	47.9	69.9	87.1	76.0	65.3	77.0	–
HRNet	256×192	N	HRNet-W32	28.5	7.1	73.4	89.5	80.7	70.2	80.1	78.9
LPN-101	256×192	N	ResNet-101	5.3	1.4	70.4	88.6	78.1	67.2	77.2	76.2
MobileNetV2	256×192	N	MobileNetV2	9.6	1.4	64.6	87.4	72.3	61.1	71.2	70.7
ShuffleNetV2	256×192	N	ShuffleNetV2	7.6	1.2	59.9	85.4	66.3	56.6	66.2	66.4
Small HRNet	256×192	N	HRNet-W18	1.3	0.5	55.2	83.7	62.4	52.3	61.0	62.1
LiteHRNet	256×192	N	LiteHRNet-18	1.8	0.3	67.2	88.0	75.0	64.3	73.1	73.3
DiteHRNet	256×192	N	DiteHRNet-18	1.8	0.3	68.3	88.2	76.2	65.5	74.1	74.2
HRFormer-T	256×192	N	HRFormer-T	2.5	1.5	70.9	89.0	78.4	67.2	77.8	76.6
LDHNet	256×192	N	LDHNet	2.2	1.4	70.6	90.5	78.2	67.8	74.8	73.7

注：“–”表示暂无数据，Y和N表示是否使用预训练。

说明LDHNet在面对陌生场景时仍能表现出稳定的模型性能。

在MPII数据集上对本文方法进行了验证，结果如表4所示，表中包含了MPII数据集中的7个人体关节点以及平均检测准确度，与LiteHRNet相比，本文方法的平均检测准确度提升了1.5%，在与DiteHRNet的对比中平均检测准确度也有着0.9%的提升，与LPN-50相比，本文方法的平均检测准确度提升了2.5%，在与一些大型网络诸如SHN、SBL-50的对比中，LDHNet不仅在参数量和计算量上有极大优势，并且平均检测准确度也有所提升，与HRNet相比，仅用不足1/10的参数量与1/5的计算量就实现了相近水准的精度表现。

综合以上的实验结果，本文LDHNet在轻量级姿态估计中具有较大的性能优势，并且模型的参数量和计算量都与当下先进的方法处于同等水准甚至更低，在与大型网络的对比中，在模型轻量化方面具有极大优势的同时，达到了相近甚至更高的性能水准，由这些结论可以得出，LDHNet有效实现了模型轻量化的目标，使模型规模与模型性能之间的关系达到了平衡。

表3　COCO测试集实验结果对比

Table 3　Comparison of experimental results of COCO test-dev set

方法	输入尺寸/像素	参数量/M	GFlops	AP/%	AP^{50}/%	AP^{75}/%	AP^{M}/%	AP^{L}/%	AR/%
CPN	384×288	–	–	72.1	91.4	80.0	68.7	77.2	78.5
SBL-152	384×288	68.6	35.6	73.0	91.7	80.9	69.5	78.1	79.0
HRNet	384×288	28.5	16.0	74.9	92.5	82.8	71.3	80.9	80.1
LiteHRNet	384×288	1.8	0.7	69.7	90.7	77.5	66.9	75.0	75.4
HRFormer-S	384×288	7.8	6.2	74.5	92.3	82.1	70.7	80.6	79.8
LDHNet	384×288	2.2	3.3	70.5	90.6	77.6	68.9	75.8	76.6

注：“–”表示暂无数据。

表4　MPII验证集实验结果对比（PCKh@0.5）

Table 4　Comparison of experimental results of MPII val set (PCKh@0.5)

方法	输入尺寸/像素	参数量/M	GFlops	Mean/%	头部/%	肩膀/%	手肘/%	腕部/%	髋部/%	膝盖/%	脚踝/%
SHN	256×256	25.1	19.1	87.5	96.5	95.3	88.4	83.5	87.1	83.5	78.3
SBL-50	256×256	34.0	12.0	88.5	96.4	95.3	89.0	83.2	88.4	84.0	79.6
HRNet	256×256	28.5	9.5	90.3	97.1	95.9	90.3	86.4	89.1	87.1	83.3
LPN-50	256×256	2.9	1.5	86.6	96.0	94.4	86.4	80.0	87.3	81.4	76.3
LiteHRNet	256×256	1.8	0.4	87.0	–	–	–	–	–	–	–
DiteHRNet	256×256	1.8	0.4	87.6	–	–	–	–	–	–	–
LDHNet	256×256	2.5	1.8	88.5	96.8	95.3	88.8	83.4	88.3	83.7	79.5

注：“–”表示暂无数据。

2.4　消融实验

在MPII数据集上进行消融实验，对Stem模块进行了消融实验以验证其作用，使用HRNet原数据预处理模块即两组3×3卷积串联与Stem模块进行对比，仅对模块进行替换其余设置均相同，结果如表5所示。

由表5结果可以看出，在使用Stem模块后，对模型的参数量和计算量都没有产生影响，但是却较大幅度地提升了模型的性能，得益于Stem模块双分支的结构，才能更大程度地保留原始的特征信息，这在较深的网络模型中显得尤为重要。

随后，针对所提出的GMNeck、GWConv与原HRNet中的残差网络模块进行消融实验，结果如表6所示。其中，对GMCneck和GWConv分别进行了单

表5　Stem模块消融实验

Table 5　Stem module ablation experiment

方法	参数量/M	Stem	GFlops	Mean/%
LDHNet	2.5	×	1.8	87.7
LDHNet	2.5	√	1.8	88.5

独的和同时的消融实验。可以看出，GMCneck 虽然在模型的轻量化方面贡献不大，但是对检测精度的提升有着较为重要的作用，而 GWConv 则对模型的轻量化发挥着重要的作用，将模型的参数量降低了一个数量级，受限于轻量级模型的有限资源检测精度受到了一定降低，对模型的轻量化作为本文的设计目标，在模型规模的缩减相比于检测精度损失不大的情况下，两个主要模块有效实现了轻量化的设计目标，使模型规模和性能达到了较好的平衡。

表 6　主要模块消融实验

Table 6　Mainly module ablation experiment

方法	参数量/M	GMCneck	GWConv	GFlops	Mean/%
LDHNet	2.5	√	√	1.8	88.5
LDHNet	2.5	×	√	1.8	88.1
LDHNet	14.7	√	×	5.5	89.6
LDHNet	14.7	×	×	5.5	89.1

此外，在 GWConv 中，本文设计了一个交互的多尺度特征提取部分，对此部分本文也单独做了消融实验，与之对比的是采用普通 3 × 3 的逐通道卷积，结果如表 7 所示。

对于轻量级模型而言，通过提取更大范围的特征信息，能够有效提升模型的性能，将多个尺度的特征信息相互融合，是一个非常有效的手段。

最后，对模型的密集连接部分进行了消融实验，

表 7　多尺度消融实验

Table 7　Multi-scale ablation experiment

方法	参数量/M	多尺度特征提取	GFlops	Mean/%
LDHNet	2.5	√	1.8	88.5
LDHNet	2.5	×	1.8	87.7

将其与直接的串行结构进行对比，结果如表 8 所示。

使用密集连接的结构使得模型的复杂度较大幅度提高，但仍处于足够轻量化的范围内。与此同时，对中间层的特征图进行充分的复用也使得模型的性能得到了大幅度的提升，证明了密集连接的结构作为本文模型的基础框架发挥了至关重要的作用。

表 8　密集连接结构消融实验

Table 8　Dense connected structure ablation experiment

方法	参数量/M	GWblock	GFlops	Mean/%
LDHNet	2.5	√	1.8	88.5
LDHNet	1.2	×	1.7	82.9

2.5　推理速度

本文选取部分方法与 LDHNet 进行了推理速度的对比测试，对于轻量级模型的测试更强调在资源有限的情况下进行，所以本文的测试环境仅使用 CPU 进行，型号为 I5-10400F，在一致的实验条件下进行测试，并在一段时间内取帧率的范围值作为结果，结果如图 7 所示，图中圆形的大小代表模型的参

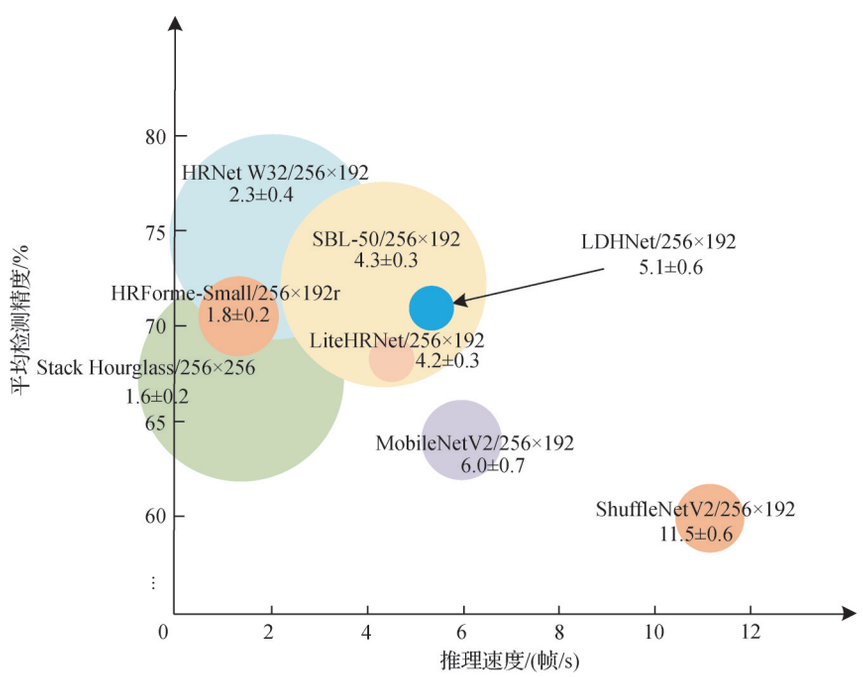

图 7　推理速度

Fig. 7　Inference speed

数量与计算量的大小，即模型规模。LDHNet相比于基线HRNet，在模型规模大幅缩减的同时推理速度也得到了较大提升，与同样作为轻量级姿态估计模型的LiteHRNet相比，速度方面仍有着较大优势。另外，对于轻量级模型而言，不能够仅仅关注于速度，同时要兼顾检测的精度，速度与检测精度都是至关重要的指标，相比于MobileNetV2和ShuffleNetV2的高速但是低精度，LDHNet更加平衡，对推理速度的验证证明了LDHNet模型的可靠性。

2.6 可视化

最后，将LDHNet对COCO2017数据集的预测结果进行可视化处理，图8为部分样本的可视化结果，在面对低光照、模糊和部分遮挡的复杂情况时，模型能够保持良好的性能表现，当面对多人的复杂场景时，模型仍旧能够保持良好的性能表现。综合这些结果可以看出，本文提出的LDHNet具有优秀的稳定性，在面对各类场景时都能够充分发挥良好性能。

图8 姿态估计可视化

Fig. 8 Visualization of human pose estimation

3 结 论

本文工作围绕着设计一个轻量级的人体姿态估计模型，面对如何在极为有限的资源限制下达到更高的检测精度这一问题，深入研究了领域内经典的以及最新的方法，选择了强劲的HRNet作为基线进行改进，通过融合密集连接网络和多尺度特征提取，实现了在有限资源下对特征图信息的充分榨取，高效地完成了特征提取的任务。另外本文还将HRNet中原本的跨分支特征交互部分进行了改动，同样对模型的轻量化和提升检测性能做出了较大贡献，通过对模型在公开数据集MPII和COCO val以及COCO test-dev数据集上的验证，本文提出的LDHNet在轻量级人体姿态估计中取得了优异的性能表现，模型的参数量相比之下也有优势，更加轻量。

除此之外，轻量级人体姿态估计的意义在于能够应用于更多设备实现实时的检测工作，通过在同样硬件环境下对模型进行推理速度的测试，验证了LDHNet在推理速度方面也有较大提升，本文仅在GPU加速环境下测试了推理速度，LDHNet能够充分利用硬件资源发挥检测性能，与Lite-HRNet相比具有更高的适用性。

本文工作仍有不足之处，模型的推理速度提升没有达到与其轻量化程度相匹配的提升，后续工作重点将在提升模型的推理速度，将模型的训练与推理相互剥离，融合重参数思想来进一步提升模型的综合性能。

参考文献(References)

Andriluka M, Pishchulin L, Gehler P and Schiele B. 2014. 2D human pose estimation: new benchmark and state of the art analysis//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE: 3686-3693 [DOI: 10.1109/CVPR.2014.471]

Cai Z D, Ying N, Guo C S, Guo R and Yang P. 2021. Research on mul-

tiperson pose estimation combined with YOLOv3 pruning model. Journal of Image and Graphics, 26(4): 837-846 (蔡哲栋, 应娜, 郭春生, 郭锐, 杨鹏. 2021. YOLOv3剪枝模型的多人姿态估计. 中国图象图形学报, 26(4): 837-846) [DOI: 10.11834/jig.200138]

Chen Y L, Wang Z C, Peng Y X, Zhang Z Q, Yu G and Sun J. 2018. Cascaded pyramid network for multi-person pose estimation//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 7103-7112 [DOI: 10.1109/CVPR.2018.00742]

Cheng B W, Xiao B, Wang J D, Shi H H, Huang T S and Zhang L. 2020. HigherHRNet: scale-aware representation learning for bottom-up human pose estimation//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 5385-5394 [DOI: 10.1109/CVPR42600.2020.00543]

Deng Y N, Luo J X and Jin F L. 2019. Overview of human pose estimation methods based on deep learning. Computer Engineering and Applications, 55(19): 22-42 (邓益侬, 罗健欣, 金凤林. 2019. 基于深度学习的人体姿态估计方法综述. 计算机工程与应用, 55(19): 22-42) [DOI: 10.3778/j.issn.1002-8331.1906-0113]

Gao K, Li W G, Shu Y, Wang Z G and Ge Y K. 2022. Multi-scale lightweight human pose estimation with dense connections. Computer Engineering and Applications, 58(24): 196-204 (高坤, 李汪根, 束阳, 王志格, 葛英奎. 2022. 融入密集连接的多尺度轻量级人体姿态估计. 计算机工程与应用, 58(24): 196-204) [DOI: 10.3778/j.issn.1002-8331.2208-0209]

Hou Q B, Zhou D Q and Feng J S. 2021. Coordinate attention for efficient mobile network design//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 13708-13717 [DOI: 10.1109/CVPR46437.2021.01350]

Howard A G, Zhu M L, Chen B, Kalenichenko D, Wang W J, Weyand T, Andreetto M and Adam H. 2017. MobileNets: efficient convolutional neural networks for mobile vision applications [EB/OL]. [2023-05-31]. https://arxiv.org/pdf/1704.04861.pdf

Huang G, Liu Z, Van Der Maaten L and Weinberger K Q. 2017. Densely connected convolutional networks//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 2261-2269 [DOI: 10.1109/CVPR.2017.243]

Iandola F N, Han S, Moskewicz M W, Ashraf K, Dally W J and Keutzer K. 2016. SqueezeNet: AlexNet-level accuracy with 50 × fewer parameters and < 0.5 MB model size [EB/OL]. [2023-05-31]. https://arxiv.org/pdf/1602.07360.pdf

Kong Y H, Qin Y F and Zhang K. 2023. Deep learning based two-dimension human pose estimation: a critical analysis. Journal of Image and Graphics, 28(7): 1965-1989 (孔英会, 秦胤峰, 张珂. 2023. 深度学习二维人体姿态估计方法综述. 中国图象图形学报, 28(7): 1965-1989) [DOI: 10.11834/jig.220436]

Lee Y, Hwang J W, Lee S, Baw Y and Park J. 2019. An energy and GPU-computation efficient backbone network for real-time object detection//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Long Beach, USA: IEEE: 752-760 [DOI: 10.1109/CVPRW.2019.00103]

Li Q, Zhang Z Y, Xiao F, Zhang F and Bhanu B. 2022. Dite-HRNet: dynamic lightweight high-resolution network for human pose estimation [EB/OL]. [2023-05-31]. https://arxiv.org/pdf/2204.10762.pdf

Lin T Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P and Zitnick C L. 2014. Microsoft COCO: common objects in context//Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland: Springer: 740-755 [DOI: 10.1007/978-3-319-10602-1_48]

Luvizon D C, Picard D and Tabia H. 2018. 2D/3D pose estimation and action recognition using multitask deep learning//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 5137-5146 [DOI: 10.1109/CVPR.2018.00539]

Ma N N, Zhang X Y, Zheng H T and Sun J. 2018. ShuffleNet V2: practical guidelines for efficient CNN architecture design//Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer: 122-138 [DOI: 10.1007/978-3-030-01264-9_8]

Mehta S, Rastegari M, Shapiro L and Hajishirzi H. 2019. ESPNetv2: a light-weight, power efficient, and general purpose convolutional neural network//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 9182-9192 [DOI: 10.1109/CVPR.2019.00941]

Newell A, Yang K Y and Deng J. 2016. Stacked hourglass networks for human pose estimation//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer: 483-499 [DOI: 10.1007/978-3-319-46484-8_29]

Sandler M, Howard A, Zhu M L, Zhmoginov A and Chen L C. 2018. MobileNetV2: inverted residuals and linear bottlenecks//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 4510-4520 [DOI: 10.1109/CVPR.2018.00474]

Sun K, Xiao B, Liu D and Wang J D. 2019. Deep high-resolution representation learning for human pose estimation//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 5686-5696 [DOI: 10.1109/CVPR.2019.00584]

Wang J D, Sun K, Cheng T H, Jiang B R, Deng C R, Zhao Y, Liu D, Mu Y D, Tan M K, Wang X G, Liu W Y and Xiao B. 2021. Deep high-resolution representation learning for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(10): 3349-3364 [DOI: 10.1109/TPAMI.2020.2983686]

Wang Q L, Wu B G, Zhu P F, Li P H, Zuo W M and Hu Q H. 2020.

ECA-Net: efficient channel attention for deep convolutional neural networks//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 11531-11539 [DOI: 10.1109/CVPR42600.2020.01155]

Wang R J, Li X and Ling C X. 2018. Pelee: a real-time object detection system on mobile devices//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal, Canada: Curran Associates Inc.: 1967-1976

Xiao B, Wu H P and Wei Y C. 2018. Simple baselines for human pose estimation and tracking//Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer: 472-487 [DOI: 10.1007/978-3-030-01231-1_29]

Xu Y, Zhang J, Zhang Q M and Tao D C. 2022. Vitpose: simple vision Transformer baselines for human pose estimation. Advances in Neural Information Processing Systems, 2022, 35: 38571-38584

Yu C Q, Xiao B, Gao C X, Yuan L, Zhang L, Sang N and Wang J D. 2021. Lite-HRNet: a lightweight high-resolution network//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 10435-10445 [DOI: 10.1109/CVPR46437.2021.01030]

Yuan Y H, Fu R, Huang L, Lin W H, Zhang C, Chen X L and Wang J D. 2021. HRFormer: high-resolution transformer for dense prediction [EB/OL]. [2023-05-31]. https://arxiv.org/pdf/2110.09408.pdf

Zhang Z, Tang J and Wu G S. 2020. Simple and lightweight human pose estimation [EB/OL]. [2023-05-31]. https://arxiv.org/pdf/1911.10346.pdf

作者简介

高坤，男，硕士研究生，主要研究方向为深度学习和人体姿态估计。E-mail：1040980522@qq.com

李汪根，通信作者，男，教授，主要研究方向为生物计算和智能计算。E-mail：1346858911@qq.com

束阳，男，硕士研究生，主要研究方向为骨骼动作识别和深度学习。E-mail：1535282558@qq.com

葛英奎，男，硕士研究生，主要研究方向为深度学习和人体姿态估计，E-mail：1013366766@qq.com

王志格，男，硕士研究生，主要研究方向为深度学习和骨骼的动作识别，E-mail：1076797329@qq.com