第 59 卷 第 18 期/2022 年 9 月/激光与光电子学进展 研究论文

激光与光电子学进展 先进成像

轻量级高分辨率人体姿态估计研究

渠涵冰，贾振堂*

上海电力大学电子与信息工程学院，上海 200090

摘要 人体姿态估计通常使用高分辨率表示的方法来实现关键点的检测，但网络参数量较大，运算较为复杂。基于此，提出了一种轻量级高分辨率人体姿态估计算法。首先，使用稠密连接网络(DenseNet)并进行轻量化改进，提出密集连接层，使得各层之间连接更加紧密，从而降低网络的运算参数，优化网络的运算速度；其次，在降低参数且精度保持不变的情况下，在多尺度融合阶段使用上采样和反卷积模块结合的融合方式，使得输出的特征信息更加丰富，检测结果更加准确；最后，利用COCO 2017验证数据集及MPII数据集进行验证。实验结果表明，在保证准确率的情况下与其他人体姿态估计算法相比，所提算法的平均精度为74.8%，运算参数减少了63.8%，网络运算复杂度缩小了8.5%，同时也到达了实时性的效果。

关键词 图像处理；人体姿态估计；高分辨率表示；多尺度融合；轻量化；改进稠密连接网络

中图分类号 TP391.4；TP183 **文献标志码** A **DOI**：10.3788/LOP202259.1810012

Lightweight and High-Resolution Human Pose Estimation Method

Qu Hanbing, Jia Zhentang*

College of Electronics and Information Engineering, Shanghai University of Electric Power, Shanghai 200090, *China*

Abstract For human pose estimation, a high-score representation method is usually adopted for detecting key points; however, this detection is difficult to achieve because of numerous network parameters and complicated calculations. In this study, to realize a closer connection between layers and achieve an enhanced lightweight nature, the densely connected network (DenseNet) is employed and densely connected layers are proposed. The network calculation parameters are reduced while the detection accuracy is maintained, and the network computing speed is optimized. Second, a fusion method that combines upsampling and deconvolution modules in the multiscale fusion stage is proposed, facilitating more abundant output feature information and more accurate detection results more accurate. Finally, the COCO 2017 and MPII datasets are used for validating the proposed method. Experimental results show that compared with other human pose estimation algorithms, the proposed method achieves an average network accuracy of 74.8%, reduces the number of operating parameters by 63.8%, and decreases the network calculation complexity by 8.5% while ensuring the accuracy of real-time effects.

Key words image processing; human body pose estimation; high-resolution representation; multi-scale fusion; lightweight; improved densely connected network

1 引 言

人体姿态估计是目前计算机视觉重要研究方向之一，其目的是检测到目标后对人体的关键点信息进行标注。随着人工智能技术的不断发展，基于人体关键点的研究在多个领域同样得到了广泛的应用，例如人机交互[1]、行为识别[2]和步态分析[3]等。近几年的研究中，大多数研究者利用更深层次的卷积神经网络来提高检测的精度，这也使得神经网络中的参数增多，运算复杂度加大，因此如何在保持准确率的情况下降低网络模型参数及其运算复杂度，从而提高检测速度，仍是人体姿态估计研究所面临的一个问题。

传统的单人人体姿态估计主要基于树状的图结构模型[4]，该方法通过观察人体不同部位和它们之间的

收稿日期：2021-05-27；**修回日期**：2021-07-04；**录用日期**：2021-08-10

基金项目：国家自然科学基金青年科学基金(61401269)

通信作者：*462458081@qq.com

空间依赖性进行推理。图模型采用的是模板匹配，即将被检测的目标分成多个部分，且各个部分有空间约束。该模型是在人为指定下的组件进行关键点的检测，其优点是计算复杂度低、实现简单，缺点是对噪声和运动时间间隔的变化比较敏感[5]。为了对人体关键点进行更可靠的观察，卷积神经网络被用在人体姿态估计中。与传统的方法相比，卷积神经网络在精度和鲁棒性等指标上都有很好的表现。2016 年，Wei 等[6]提出了卷积姿态机结构，该方法的网络结构分为多个阶段，将上一阶段的输出融入下一阶段的输入，并对每个阶段进行监督训练，进一步提高关键点的检测效果。但在单独对每个候选框进行估计时，都不会考虑人体之间相互遮挡或截断的问题，这使得估计的关节点不能与正确的人体相关联，从而导致估计错误[7]。Newell 等[8]提出了 Stacked hourglass 网络，该网络一方面利用多分辨率的 heat map 学习关节点的局部位置特征，另一方面通过多尺度感受野机制学习并获得关节点之间的结构特征[9]，取得了一定的效果，但是该网络是从高分辨率到低分辨率再恢复到高辨率的，在此过程中所产生的误差对低分辨率图像的检测结果不佳。2017 年，文献[10]提出的 feature pyramid 通过 pyramid residual module 来解决深度模型的尺度不变性问题，并且对权重初始化提出了新的策略。然而，这些方法都只是对单人进行检测，并给出了相关人体的位置和大小。

对于多人人体姿态估计有很多种方法，大致可分为自上而下的和自下而上的。自下而上的方法先识别人体关键点，然后对检测到的关键点进行聚类组合。Cao 等[11]提出了 OpenPose 模型，该模型将特征图像输入一个由关键点检测和肢体连接并行结构的卷积网络中，再将两个网络生成的图像通过拼接得到人体姿态检测图像。而自上而下的方法先进行每个人物的识别，然后再检测每个人的关键点。Cascaded pyramid network(CPN)[12]采用级联结构，通过一个类似特征金字塔的 GlobalNet 和一个用来连接所有金字塔作为语义信息的 RefineNet 来获取关键点的位置，但是只采用低分辨率的信息，而未使用高分辨率的信息。Sun 等[13]提出的 HRNet 利用高分辨率的特征，整个网络始终存在高分辨率，该网络从高分辨率子网开始为第一阶段，并且将该子网分为 4 个具有不同分辨率的并行子网，在一定程度上提高了检测的准确度。然而该网络存在一定的缺陷，特征图只通过最紧邻上采样得到，图像质量有待提高，计算复杂且实时性较差。2020 年，Cheng 等[14]在 HRNet 的基础上提出了 HigherHRNet，HigherHRNet 在 HRNet 的末端进一步使用高分辨率的方法，取得了更好的效果。但是由于采用了保持高分辨率的方法，该网络在提高准确率的同时也提高了网络的复杂度，增加了网络运算量。

针对以上问题，本文提出了一种轻量级高分辨率人体姿态估计算法。首先，运用和 HRNet 相似的实例化主干，引入稠密卷积网络(DenseNet)，并对其加以改进。在原始 DenseNet 的卷积结构上加以变换，在网络结构的 3×3 卷积前加一个 1×1 卷积组成瓶颈结构，这样整个网络的运算参数量也随之减少，并提出一种提高准确度且减少网络参数的稠密集残差网络。然后，在高分辨率最后融合阶段采用逐层上采样的方法，将低分辨率的信息转换为高分辨率，从而提取到更多的信息，得到高分辨率的输出，再结合反卷积模块得到更高质量的特征图输出。最后，利用 COCO 数据集和 MPII 数据集将所提算法与其他人体姿态估计算法进行了准确度和运算量的实验对比。

2 相关工作

2.1 多尺度融合

多尺度融合最直接的方法就是将多个分辨率图像分别接入多个网络，然后融合特征图并输出。Hourglass[8]中所采用的多尺度融合方法是通过跳跃连接，逐步将下采样得到的低分辨率图和上采样得到的高分辨率图结合起来，从高分辨率到低分辨率和低分辨率到高分辨率这个过程是对称的。HRNet 第 1 阶段以高分辨率作为子网，后阶段将该子网分为 4 个具有不同分辨率的并行子网，通过并行连接高分辨率到低分辨率，从而确保每个阶段都有高分辨率特征图，并通过反复融合由高到低子网产生的多分辨率表示来生成可靠的高分辨率表示。融合操作之后，所有分辨率下得出的信息都将用于输出最终的高分辨率热图。因此，本实验组以 HRNet 作为基本模块。

2.2 特征金字塔

金字塔状的表示在目标检测和物体分割框架已被广泛应用于处理多种尺度变化。特征金字塔网络[15]使用自上而下上的方式将主干网络进行扩展，该方法使用双线性上采样和横向连接逐渐将特征分辨率从 1/32 恢复到 1/4。然而，这种金字塔形表示在自下而上的多人姿势估计中很少得到应用。HRNet 是一个基于高分辨率的特征金字塔，它将金字塔扩展到不同的方向，从 1/4 分辨率的特征开始，并生成具有更高分辨率的特征金字塔。

2.3 稠密连接网络

稠密卷积连接网络(DenseNet)[16]使用一种将网络提炼为简单连接模式的体系结构。为了确保网络中各层之间的信息流最大化，稠密卷积连接网络将所有层(具有匹配的特征图大小)直接相互连接。为了保留前馈性质，每个层都从所有先前的图层获取附加输入，并将其自身的特征图传递给所有后续层。与 ResNet 相比，DenseNet 通过串联功能来组合特征，而不是在要素传递到图层之前通过求和进行组合。

3　网络结构

3.1　网络模型

本实验组采用与HRNet相似的方式作为实例化主干。所提网络一共有4个阶段且并行连接，使特征图分辨率降为输入原始图片的1/4，各层分辨率依次减小1/2，对应的通道数是上一阶段的2倍。图像的初始分辨率为192×256，首先使用两个3×3的卷积将图像分辨率降低到原始图像的1/4，即第1阶段由分辨率为48×64的子网组成，第2阶段由分辨率为48×64和24×32的子网并联组成，第3阶段由分辨率为48×64、24×32和12×16的子网并联组成，第4阶段由分辨率为48×64、24×32、12×16和6×8的子网并联组成。第1阶段由四层稠密卷积单元构成，将特征图的宽度缩小为32。在第2、第3、第4阶段HRNet采用的多分辨率模块数分别为1、4、3个，而本实验组在这3个阶段使用1、3、2个多分辨率块，每个模块都有4个稠密连接网络，每个分辨率中的稠密残差单元都由两个卷积构成。因此，在与HRNet模型比较时，所提模型也有更多的层次，网络的深度也没有减少。在第4阶段的最后一个模块中，对后三个阶段的特征图使用由下向上逐级采样的融合方法与来自第1子网的特征图进行融合。图1为所提网络整体结构，其中每个特征图模块包含4个卷积操作特征图，R代表输入图像的分辨率。

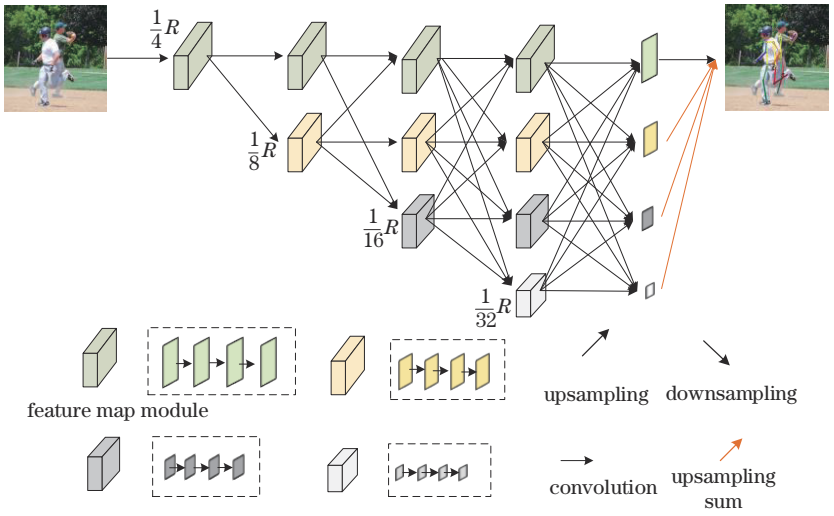

图1　网络结构

Fig. 1　Network structure

3.2　改进型稠密连接网络

随着网络的加深，运算参数也会随之增大，为了解决这一问题，本实验组使用稠密卷积连接网络。稠密卷积网络(DenseNet)是一种更密集的连接方式，在同一个Dense block中要求特征图保持相同的大小，利用前向传播方式，将每一层与其余层密集连接，能够充分利用前后层特征信息，从而获取更为丰富的特征图，具体如图2所示。第n层的输入不仅与$n-1$层的输出相关，还与之前所有层都有关，可描述为

$$\chi_n = H_n\left[\chi_0, \chi_1, \cdots, \chi_{n-1}\right], \tag{1}$$

式中：[ ]代表拼接，即将χ_0到χ_{n-1}层的所有输出的特征图按通道数组合；H_n为第n层的非线性变换。

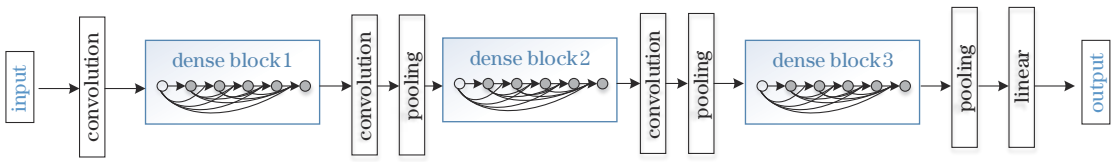

图2　DenseNet

Fig. 2　DenseNet

人体姿态估计利用算法将图片中人的所有关节找出(如肩部、手腕、膝盖等)。对于一个检测K个关节点的位置，有$P_k \in M$，其中P_k代表第k个关节点的位置，M代表一张分辨率为$w \times h$的图片的所有位置。改进型的高分表率姿态估计方法利用轻量级主干对图片进行特征信息的提取，并在最后融合阶段加强高质量特征图的输出。与ResNet相比，稠密卷积具有更少的网络参数，精度与ResNet相同，而且也能更好地解决过度拟合问题。原始的HRNet在主干网络中使用的是残差块，在不损失精度且减少网络参数量的情况下，所提网络采用的密集连接模块如图3所示。在密集单元中不同于残差网络中直接进行上下层输出的相加，而

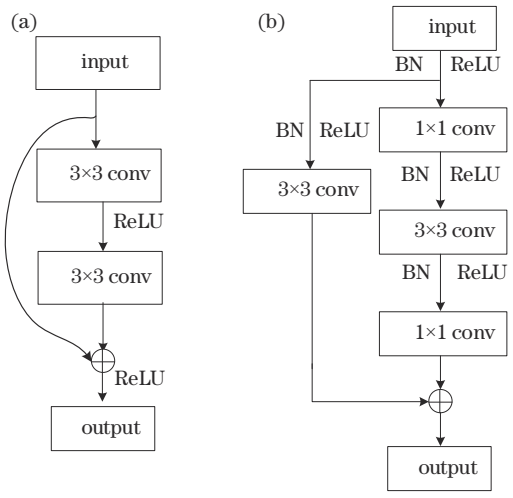

图 3　改进型稠密网络模块。(a)残差单元;(b)密集单元

Fig. 3　Improved dense network module. (a) Residual unit; (b) dense cell

是进行通道维度上的连接,这样可以使上层的输出直接和下一层的输出连接起来,使得特征重用及梯度的快速传递,卷积层输入的表达式为

$$I_n = k_0 + (n-1)k, \tag{2}$$

式中:I_n为各层的输入;k_0为最开始的输入特征图的通道数;k为网络通道增长率。模块后面输入的特征包含前层的输出特征,卷积层输出特征的高频复用,使不同层次的特征得以更好融合利用;同时跨连接也可以避免梯度消失问题,为了防止大量特征拼接导致计算成本过高,任一层的输出特征都通过降维得到较小的通道维度。

在原始 DenseNet 的卷积结构上加以变换,在网络结构的 3×3 卷积前加一个 1×1 卷积组成瓶颈结构,这样整个网络的运算参数量也随之减少;为了防止在对网络进行训练的实验过程中产生较大的损失波动,在每个密集层单元中增加一个 3×3 的卷积,让不同密集层直接连接更加快捷,最终使得损失达到一个较低水平。每个密集层之间都有密集连接,并且每个层都将先前所有的特征图作为输入。并将原始 HRNet 中的所有残差单元使用以上的密集层连接网络代替,实验结果表明,参数的运算量大幅减少。

3.3　融合方式

在多尺度融合中,采用并行子网络引入交换单元,使每个子网络重复地从其他并行子网络接收信息。将第 3 阶段划分为几个交换块,每个块由 3 个并行卷积单元和一个跨越并行单元的交换单元组成。不同分辨率的特征图进行多尺度的融合时,采用步长为 3×3 的卷积进行下采样,步长为 1×1 的卷积进行最近邻上采样,多尺度融合可描述为

$$\begin{matrix} C_{31}^{1}\searrow & & \nearrow C_{31}^{2}\searrow & & \nearrow C_{31}^{3}\searrow \\ C_{32}^{1}\rightarrow & \varepsilon_3^2 & \rightarrow C_{32}^{2}\rightarrow & \varepsilon_3^2 & \rightarrow C_{32}^{3}\rightarrow \\ C_{33}^{1}\nearrow & & \searrow C_{33}^{1}\nearrow & & \searrow C_{33}^{3}\nearrow \end{matrix} \varepsilon_3^3, \tag{3}$$

式中:C_{sr}^{b}代表在阶段 s 中第 b 块分辨率为 r 的卷积单元;ε_s^b为相应的交换单元。

四阶分辨率表示将在融合之前的最后阶段的最后一个块中输出,因此采用怎样的融合方式是一个重要的问题,对整个网络的性能都有影响。原始 HRNet 中的输出只是对高分辨率的表示,忽略了其他三个表示形式,如[图 4(a)]所示。本实验组将原始 HRNet 中的最后阶段的输出融合方式加以变换,在第 4 阶段后输出的各个分辨率特征图后,采用从下至上即低分辨率到高分辨率进行上采样的方式融合,并将得到的上采样结果加入最后一个分辨率的特征中,从低分辨率到高分辨率这样逐次循环,直到输出最高分辨率的特征。这样的融合方式可得到不同层次的信息,使得输出的特征信息更加丰富,检测结果更加准确。受到 HigherHRNet 的启发,在上述的融合阶段输出的特征图后,利用起始阶段生成的 1/4 分辨率图形和原来的融合阶段输出的特征图进行反卷积操作。反卷积模块以上两种图作为输入,可以生成质量更高的特征图并生成分辨率比输入特征图大 2 倍的新特征图,改进融合方式如[图 4(b)]所示。

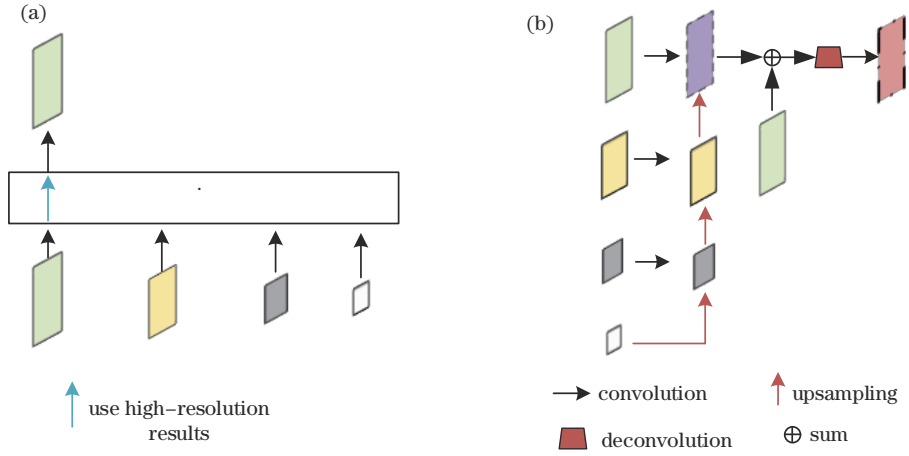

图 4　融合方式对比。(a)原 HRNet 的最后融合方式;(b)改进的最后融合方式

Fig. 4　Comparison of fusion methods. (a) Final fusion method of original HRNet; (b) improved final fusion method

4 实验及结果分析

使用Pytorch深度学习框架作为实验框架，软件操作环境为Windows 10系统，使用Python3.6作为编程语言。CPU为i7-9700K，GPU为RTX2080ti，显存为28 GB。以1×10^{-3}作为实验的初始学习率，周期120后将学习率调整为1×10^{-4}，总的训练周期为210，批训练数量设置为32。使用Adam方法对训练进行优化。

4.1 实验数据及评价标准

4.1.1 实验数据

使用两种数据集进行验证，一个是MPII数据集，它是从大量的实际活动中获取的图片，并且对全身的姿态都有所注释。MPII数据集中有大约25000张图像，这些图片包含有40000个注释关节的人，其中12000张图像用于测试，其余的则用于验证。另一个为COCO数据集[17]，MS COCO是由微软公司提供的用于物体检测、关键点检测的数据集。本实验组将COCO2017中118287张训练集图片作为训练样本，5000张验证集作为验证样本，全身18个人体关键点作为检测范围，这些关键点有可见、不可见和不在图上3种状态。COCO数据集中的关键点标注顺序如表1所示。

表1 COCO数据集中的关键点

Table 1 Key points in COCO dataset

Serial number	Name	Serial number	Name
0	Nose	9	Right knee
1	Neck	10	Right ankle
2	Right shoulder	11	Left hip
3	Right elbow	12	Left knee
4	Right wrist	13	Left ankle
5	Left shoulder	14	Right eye
6	Left elbow	15	Left eye
7	Left wrist	16	Right ear
8	Right crotch	17	Left ear

4.1.2 评价标准

为了计算真实值和预测的人体关键点的相似程度，对于COCO数据集，依据COCO数据集官网所提供的评估标准object keypoint similarity (OKS)作为评价模型的标准，OKS在0～1之间取值，其值越接近1表明预测效果越好，OKS的表达式为

$$S_{\mathrm{OKS}}=\frac{\sum_i \frac{-d_i^2}{\exp\left(2s^2k_i^2\right)}\cdot\delta(v_i>0)}{\sum_i\left[\delta(v_i>0)\right]},\tag{4}$$

式中：d_i是每个真实关键点与预测关键点之间的欧几里得距离；v_i是关键点的可见性标；s是物体尺度；k_i是每个关键点的控制衰减常数。

使用标准的平均准确率和召回率对实验结果进行验证：AP(OKS在0.50，0.55，…，0.90，0.95这10个位置准确度的平均值)、AP^{50}(OKS在0.5处的准确度)、AP^{75}(OKS在0.75处的准确度)、AP^{L}(大型物体准确度)、AP^{M}(中等物体准确度)、AR(OKS在0.50，0.55，…，0.90，0.95这10个位置的平均召回率)。还利用了浮点运算数(FLOPs)[18]，FLOPs的表达式为

$$Q_{\mathrm{FLOPs}}=2HW\left(C_{\mathrm{in}}K^2+1\right)C_{\mathrm{out}},\tag{5}$$

式中：HW为输入特征图大小；C_{in}为输入通道数；K为卷积核大小；C_{out}为输出通道数。FLOPs通常理解为计算量，可以用来衡量算法/模型的复杂度。

对于MPII数据集，采用Head-normalized Probability of Correct Keypoint(PCKh)[19]作为评估模型的方法。如果关节落在groundtruth位置的αl个像素内，则关节是正确的，其中α是一个常数，l为头部尺寸对应于groundtruth头部边界的对角线长度的60%。这里使用各个关节点的准确度及PCKh@0.5(@=0.5)作为评估参数。

4.2 模型训练

在训练阶段，数据扩充采用与HRNet相同的方式，即随即旋转$(-45°,45°)$，随即缩放(0.65，1.35)，并将输入图像调整到高宽之比为4∶3的比例。在对COCO数据集进行训练时设置了两个阶段，第1阶段是0～120周期，设置学习率为1×10^{-3}，可以使损失值在较快的时间内达到一个较低水平；第2阶段的学习率为1×10^{-4}，使损失值平稳减小到最低，直至相对稳定。损失函数在训练过程的变化图如图5所示。

损失值在第1阶段前段下降较快，当迭代次数达到40时，下降变缓并逐渐趋于平稳，如[图5(a)]所示。在第2阶段，当迭代次数为140时，趋于稳定，最后在0.00035左右收敛并得到网络的权重，如[图5(b)]所示。

4.3 实验结果

将CPN模型、SimpleBaseline-50模型[20]、SimpleBaseline-101模型[20]、HRNet模型、HRNetV1模型[13]、HigherHRNet模型和所提模型这7种模型进行对比，结果如表2所示。在网络的参数量方面，HRNet-32模型与SimpleBaseline-50模型、HigherHRNet模型和SimpleBaseline-101模型相比分别缩小5.5，0.1，24.5 MB，而HRNet模型是所提改进模型的2.8倍；在运算复杂度上，所提模型比原始HRNet缩小了8.4%，比HigherHRNet缩小了86.4%；在模型大小上，HRNet模型比CPN缩小将近65.3%，比SimpleBaseline-50缩小15.5%，而所提模型比HRNet模型减少了71.9%。所提模型在COCO数据集下检测的准确度与HRNet的精确度基本相同，但在AP^{50}处的准确率高出HRNet模型2个百分点。在MPII数据集下测试了七个部位的准确率，结果如表3和图6所示。相比HRNet模型，准确率有所降低，但保持在一定的范围内且减少了计算量。从图6可以直观地看出，所提方法在各个部位检测的准确度上和其他方法基本一致，模型得以优化。从表2、

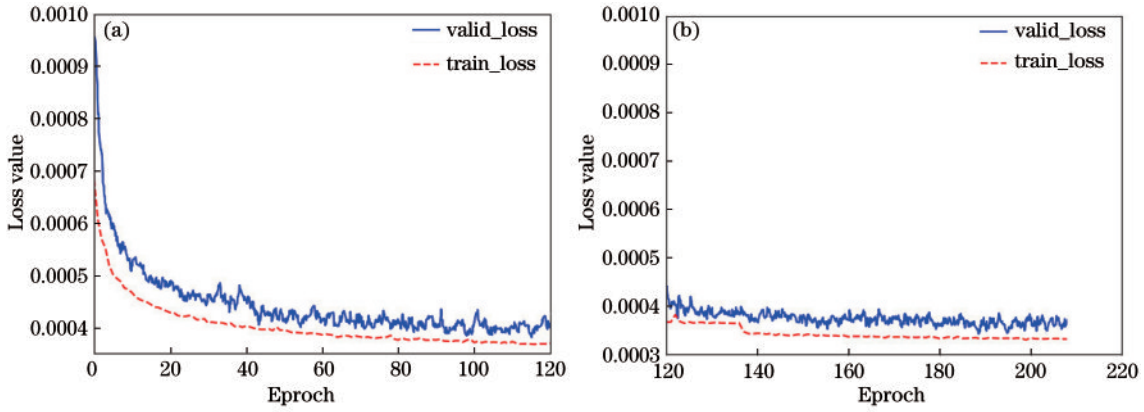

图 5　模型训练损失函数曲线。(a)第 1 阶段损失值；(b)第 2 阶段损失值

Fig. 5　Model training loss function curve. (a) First stage loss value; (b) second stage loss value

表 2　COCO 数据集下不同方法的验证比较

Table 2　Validation and comparison of different methods under COCO dataset

Method	#Params /MB	GFLOPs	Model size /MB	AP /%	AP^{50} /%	AP^{75} /%	AP^{L} /%	AP^{M} /%
CPN	27.0	6.2	314.0	68.6				
SimpleBaseline-50	34.0	9.0	129.0	70.4	88.6	78.3	67.1	77.2
SimpleBaseline-101	53.0	12.4	202.0	71.4	89.3	79.3	68.1	78.1
HRNetV1	28.5	16.0	109.4	74.9	92.5	82.8	71.3	80.9
HigherHRNet	28.6	47.9	109.8	66.4	87.5	72.8	61.2	74.2
HRNet	28.5	7.1	109.0	74.4	90.5	81.9	70.8	81.0
Proposed method	10.1	6.5	30.6	74.8	92.6	83.2	72.2	77.7

表 3　MPII 数据集下不同方法的验证比较

Table 3　Comparison of different methods under MPII dataset　　unit: %

Method	Head	Shoulder	Elbow	Wrist	Crotch	Lap	Ankle	Mean
DeeperCut[21]	97.2	94.5	87.3	82.4	86.2	81.7	77.2	86.6
SimpleBaseline-50	96.4	95.3	89.0	83.2	88.4	84.0	79.6	88.5
SimpleBaseline-101	96.9	95.9	89.5	84.4	88.4	84.5	80.7	89.1
HRNet	97.1	95.9	90.3	86.4	89.1	87.1	83.3	90.3
Proposed method	97.8	95.4	90.1	83.9	88.5	87.2	82.8	88.9

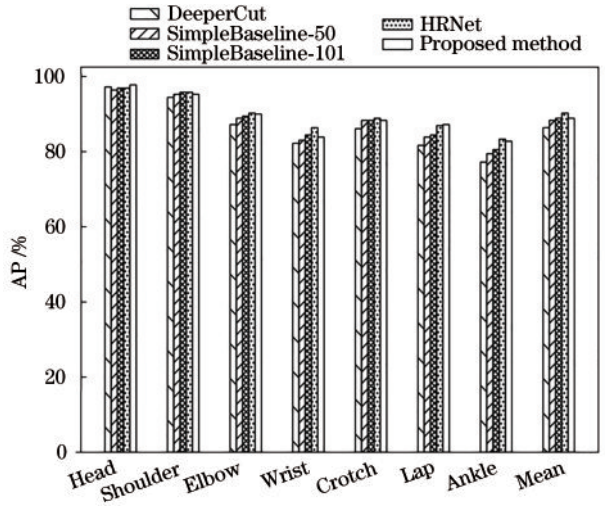

图 6　不同方法各部位预测精度比较

Fig. 6　Comparison of prediction accuracy of various parts of different methods

表 3 可以看出，在精度和模型大小之间权衡时，所提方法在保证精度的同时也缩短了模型训练的时间，降低了模型的大小。综合以上两点，所提方法在检测过程中具有较好的效果。

4.4　消融实验分析

通过改进后的密集网络和融合方式两种方法对网络模型进行优化，在平均准确率提升的情况下，减小了网络运算参数和运算复杂度，对改进前的稠密模块和改进后的稠密模块方法分别在 COCO2017 验证数据集下进行比较。结果如表 4 所示，其中 Proposed(A)代

表 4　COCO 验证数据集下消融实验

Table 4　Ablation experiment under COCO validation dataset

Model	#Params /MB	AP /%
HRNet	28.5	74.4
Proposed(A)	9.8	73.6
Proposed(B)	10.1	74.8

表改进前的稠密模块，Proposed(B)代表改进后的稠密模块的方法。

从表 4 可以看出，在未对稠密模块进行改进的实验下，网络的参数量为 9.8 MB，比 HRNet 减少了 18.7 MB，但代价是精确度减小。而改进后的方法相比于改进前在参数量几乎不变的情况下，使得检测的准确度提升了 1.2 个百分点，比 HRNet 参数量减少了 63.8% 且准确度上升了 0.4 个百分点。实验结果表明，改进后的网络使运算参数得以大幅降低，并且网络性能也有所提升。

4.5　实时性分析

将所提方法与 CPMs 和 HRNet 两种方法进行实时性分析，对比方法在 MPII 数据集下比较训练时长、单张图像的检测速度及准确度。表 5 为速度实时性比较结果。

从表 5 可以看出：所提模型在训练时长是 CPMs 的近三分之一，比 HRNet 缩短了 9 h，节省了训练时间；检测准确度方面，所提模型比 CPMs 高 3.4 个百分点，比 HRNet 低 1.4 个百分点；但单张图像检测的时间为 24 ms，检测速度最快，达到了实时的效果。

表 5　速度实时性比较

Table 5　Speed real-time comparison

Model	Training time /h	Single image detection time /ms	Accuracy (PCK) /%
CPMs	131	106.3	85.5
HRNet	54	52.2	90.3
Proposed model	45	24.0	88.9

4.6　可视化分析

采用 COCO2017 验证数据集进行可视化分析实验，随机抽取数据集中含有单人和多人的图片，得到的检测效果图如图 7 所示。当图片中的人物胳膊或腿有交叉重叠时，网络也能很好检测出各个部位的关键点位置，在[图 7(a)]中，人物紧握棒球棒使得左侧有交叉重叠，但是左手腕和左肘也能被成功检测到。在[图 7(b)]中，人物的中间有网状物或者前方有行李箱等障碍物遮挡，即使在前方有行李箱的遮挡，图片中女孩的关键点也被网络检测出来并且右膝和右脚踝能够被预测出来，说明网络有一定的抗干扰、抗噪性能。对于密集人群检测结果如[图 7(c)]所示，无论人物是正面还是背面，网络都能较好地检测和与预测关键点的位置，表明所提网络在鲁棒性与准确度方面有较高的性能。

图 7　验证结果。(a)关键部位重叠检测图；(b)障碍物遮挡检测图；(c)多人检测图

Fig.7　Validation results. (a) Key parts overlap detection map; (b) obstacle occlusion detection map; (c) multi-person detection map

4.7　算法不足

图 8 为所提模型表现不足的样例，所提基于高分辨率改进的模型也存在一些表现不足的问题：1)在图像中对于远处即较小的目标无法进行准确检测；2)在多人情况下对肢体动作的误判导致检测错误。

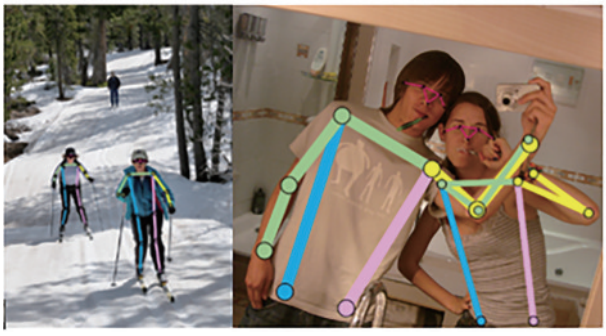

图 8　模型表现不足的样例分析

Fig. 8　Sample analysis of insufficient model performance

5　结　　论

针对利用高分辨率特征进行人体姿态估计时，网络的运算参数及其复杂度较大的问题，提出了一种密

集型轻量级网络，使得模型在运算数量及检测速度方面表现良好；同时在对单幅图片的检测中用时 24 ms，达到了实时的效果。在 COCO2017 数据集的实验结果表明，与原始的 HRNet 相比，人体关键点检测精度有所提高。如何对模型进一步优化并将轻量级的模型应用到实际工作中将是下一步的研究方向。

参 考 文 献

[1] Vuletic T, Duffy A, Hay L, et al. Systematic literature review of hand gestures used in human computer interaction interfaces[J]. International Journal of Human-Computer Studies, 2019, 129: 74-94.

[2] 徐志京, 王东. 基于双路循环生成对抗网络的多姿态人脸识别方法[J]. 光学学报, 2020, 40(19): 1910002.
XuZhijing, Dong Wang. Multi-Pose Face Recognition with Two-Cycle Generative Adversarial Network[J]. Acta Optica Sinica, 2020, 40(19): 1910002.

[3] 赵心驰, 胡岸明, 何为. 基于卷积神经网络和 XGBoost 的摔倒检测[J]. 激光与光电子学进展, 2020, 57(16): 161024.
Zhao X C, Hu A M, He W. Fall detection based on convolutional neural network and XGBoost[J]. Laser & Optoelectronics Progress, 2020, 57(16): 161024.

[4] Fischler M A, Elschlager R A. The representation and matching of pictorial structures[J]. IEEE Transactions on Computers, 1973, 22(1): 67-92.

[5] 王亮, 胡卫明, 谭铁牛. 人运动的视觉分析综述[J]. 计算机学报, 2002, 25(3): 225-237.
Wang L, Hu W M, Tan T N. A survey of visual analysis of human motion[J]. Chinese Journal of Computers, 2002, 25(3): 225-237.

[6] Wei S H, Ramakrishna V, Kanade T, et al. Convolutional pose machines[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 4724-4732.

[7] 卢健, 杨腾飞, 赵博, 等. 基于深度学习的人体姿态估计方法综述[J]. 激光与光电子学进展, 2021, 58(24): 2400005.
Lu J, Yang T F, Zhao B, et al. Review of deep learning-based human pose estimation[J]. Laser & Optoelectronics Progress, 2021, 58(24): 2400005.

[8] Newell A, Yang K Y, Deng J. Stacked hourglass networks for human pose estimation[M]//Leibe B, Matas J, Sebe N, et al. Computer vision-ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016, 9912: 483-499.

[9] 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, 55(19): 22-42.
Deng Y N, Luo J X, Jin F L. Overview of human pose estimation methods based on deep learning[J]. Computer Engineering and Applications, 2019, 55(19): 22-42.

[10] Yang W, Li S, Ouyang W L, et al. Learning feature Pyramids for human pose estimation[C]//2017 IEEE International Conference on Computer Vision, October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 1290-1299.

[11] Cao Z, Simon T, Wei S H, et al. Realtime multi-person 2D pose estimation using part affinity fields[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 1302-1310.

[12] Chen Y L, Wang Z C, Peng Y X, et al. Cascaded pyramid network for multi-person pose estimation[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 7103-7112.

[13] 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 (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 5686-5696.

[14] Cheng B W, Xiao B, Wang J D, 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), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 5385-5394.

[15] 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, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 936-944.

[16] Huang G, Liu Z, van der Maaten L, et al. Densely connected convolutional networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 2261-2269.

[17] Vinyals O, Toshev A, Bengio S, et al. Show and tell: lessons learned from the 2015 MSCOCO image captioning challenge[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 652-663.

[18] Molchanov P, Tyree S, Karras T, et al. Pruning convolutional neural networks for resource efficient inference[EB/OL]. (2016-11-19) [2021-04-06]. https://arxiv.org/abs/1611.06440.

[19] Yang Y, Ramanan D. Articulated pose estimation with flexible mixtures-of-parts[C]//CVPR 2011, June 20-25, 2011, Colorado Springs, CO, USA. New York: IEEE Press, 2011: 1385-1392.

[20] Xiao B, Wu H P, Wei Y C. Simple baselines for human pose estimation and tracking[M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11210: 472-487.

[21] Insafutdinov E, Pishchulin L, Andres B, et al. DeeperCut: a deeper, stronger, and faster multi-person pose estimation model[M]//Leibe B, Matas J, Sebe N, et al. Computer vision-ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016, 9910: 34-50.