文章编号：1006-3080(2025)03-0341-12 DOI：10.14135/j.cnki.1006-3080.20240722001

基于重参数化大核卷积的高分辨率姿态估计

陈佳艺， 黄晓宇， 吴胜昔， 王学武
（华东理工大学能源化工过程智能制造教育部重点实验室，上海 200237）

摘要：尽管人体姿态估计领域的研究已取得显著进展，但面对动态场景变化、目标遮挡及背景复杂等难题，实现高精度、强鲁棒性的姿态估计依然面临巨大挑战。为解决这些问题，特别是关键点遮挡、重合及复杂环境干扰问题，本文提出了一种融合大核卷积技术的高分辨率人体姿态估计模型(RepLK-HRNet)。该模型的核心在于特征提取网络的独特设计，通过引入重参数化大核卷积策略，增强了模型捕捉多尺度、多层次特征信息的能力，同时通过调整网络结构，显著降低了参数量和计算复杂度。实验结果表明，相较于传统的高分辨率网络(HRNet)模型，RepLK-HRNet模型在标准数据集MS COCO2017上的精度提高了1.83%，在遮挡数据集OCHuman上的精度提高了23.7%，计算复杂度参数Params和GFLOPs分别下降了63.84%、37.69%。RepLK-HRNet模型在常规及遮挡、关键点混淆等条件下的人体姿态估计精度均实现了显著提升，展现了出色的鲁棒性和泛化能力，同时还满足了实际应用中对计算效率和存储空间的要求。

关键词：姿态估计；重参数化大核卷积；HRNet；感受野；特征融合

中图分类号：TP273 **文献标志码：**A

近年来，随着深度学习技术的飞速发展，尤其是卷积神经网络 (CNN) 的广泛应用，人体姿态估计 (HPE) 领域取得了显著进展。ResNet[1] 及其变体[2-4] 等深度 CNN 模型作为该领域的主流，展示了强大的特征提取能力。在此背景下，基于沙漏网络[5-6]、级联金字塔网络 (CPNs)[7-8]以及高分辨率网络 (HRNet)[9] 等新型架构通过融合多尺度特征，显著提升了姿态估计的准确性。2020年以后，Transformer[10] 模型在视觉任务中崛起，ViTs (Vision Transformers)[11-12] 在图像分类[13-14]、语义分割[15-16] 和目标检测[17-18] 等任务上展现出的优越性能，使得 ViTs 学习视觉表征逐渐成为 CNN 的一种替代方案。TokenPose[19]、TransPose[20]、HRFormer[21]和 ViTPose[22] 从不同的角度解释了 Transformer 在姿态估计领域的优越性。

尽管 ViTs 在某些任务上表现优异，但其在处理关键点遮挡、重合及复杂环境干扰等问题上存在一定的局限性。ViTs 的图块分割方式[18] 可能破坏关键点间的连续性，增加处理难度，且对遮挡较为敏感。在处理高分辨率图像时计算复杂度较高，ViTs 可能导致延迟增加，不适合资源受限的场景[21]。相比之下，大核卷积[23-25] 通过其扩大的感受野和参数共享特性，能更有效地捕捉不同关键点之间的空间关系，减少关键点重合带来的混淆，并在遮挡情况下利用上下文信息推断被遮挡关键点的位置。相比于传统的小卷积核堆叠增加感受野方法，大卷积核可以在一定程度上减少网络层数，简化网络结构。在处理高分辨率图像或具有大尺度特征的任务时，大卷积核更具优势。除了 Inception[26-27] 等少数老式模型外，大核模型在 VGG-Net[28] 之后就不再流行。

近年来，随着计算能力的提升和深度学习理论的进步，大核卷积再次受到研究者的关注，一些新的网络架构和技术被提出，以优化大核卷积的计算性

收稿日期： 2024-07-22
基金项目： 国家自然科学基金 (62076095)
作者简介： 陈佳艺 (2001—)，女，甘肃人，硕士生，研究方向为机器学习与人工智能。E-mail: jiayichan0107@163.com
通信联系人： 吴胜昔，E-mail: wushengxi@ecust.edu.cn
引用本文： 陈佳艺，黄晓宇，吴胜昔，等. 基于重参数化大核卷积的高分辨率姿态估计 [J]. 华东理工大学学报(自然科学版)，2025，51(3)：341-352.
Citation: CHEN Jiayi, HUANG Xiaoyu, WU Shengxi, *et al*. High-Resolution Pose Estimation Based on Reparameterized Large Kernel Convolution[J]. Journal of East China University of Science and Technology, 2025, 51(3): 341-352.

能和模型效率。结构重参数化由 Ding 等[29]于 2021 年在 RepVGG 一文中首次正式提出，通过参数的等价转化实现结构的等价转换。这种方法通过将小核卷积和批量归一化(BN)运算的参数合并到大核卷积中，训练时小核卷积模拟大核卷积的效果，并在推理时将小核卷积“折叠”为大核卷积，可以在保持模型性能的同时，减少训练时的计算复杂度和优化难度。Ding 等[30] 通过结构重参数化技术实现了大核卷积的高效使用，并在后续工作中提出了参数高效的大核网络架构[31]，在用于音频、视频、点云、时间序列和图像识别等视觉任务中表现出了优异效果。2022 年，Hu 等[32] 提出了在线卷积重参数化 (OREPA)，即一个两阶段的 pipeline，旨在通过将复杂的 training-time block 压缩成单个卷积来减少巨大的训练开销。2023 年，Cai等[33] 提出了一种可重参数化的重新聚焦卷积(RefConv)，RefConv可以在不引入任何额外推理成本或改变原始模型结构的情况下，显著提高多种基于 CNN 模型的性能。

本文提出了一种结合空洞重参数化大核卷积的高分辨率人体姿态估计模型 (RepLK-HRNet)。该模型以 HRNet 网络为基础框架，提出一种多层次特征提取和多尺度特征融合相结合的特征提取方法，增强模型捕获特征的能力，在处理关键点遮挡、重合及复杂环境干扰等视觉任务时获得更丰富的语义信息。同时引入空洞重参数化卷积来增加感受野，从而在不增加参数量的前提下，有效扩大卷积层的视野范围，并结合重参数化技术优化网络结构，以减少计算复杂度和模型大小，实现了模型的轻量化设计。RepLK-HRNet 模型不仅能够保持高分辨率特征图的细节信息，还能够通过大核卷积扩大感受野，增强模型对全局信息的捕捉能力，从而在背景复杂和遮挡情况下实现更精准的人体姿态估计。

1 RepLK-HRNet 整体结构

RepLK-HRNet 的整体框架包括下采样层、主体和回归器层，如图 1 所示，其中 DR 层表示空洞重参数块，SE 层表示压缩激励(Squeeze and Excitation)，FFN 层表示前馈网络。

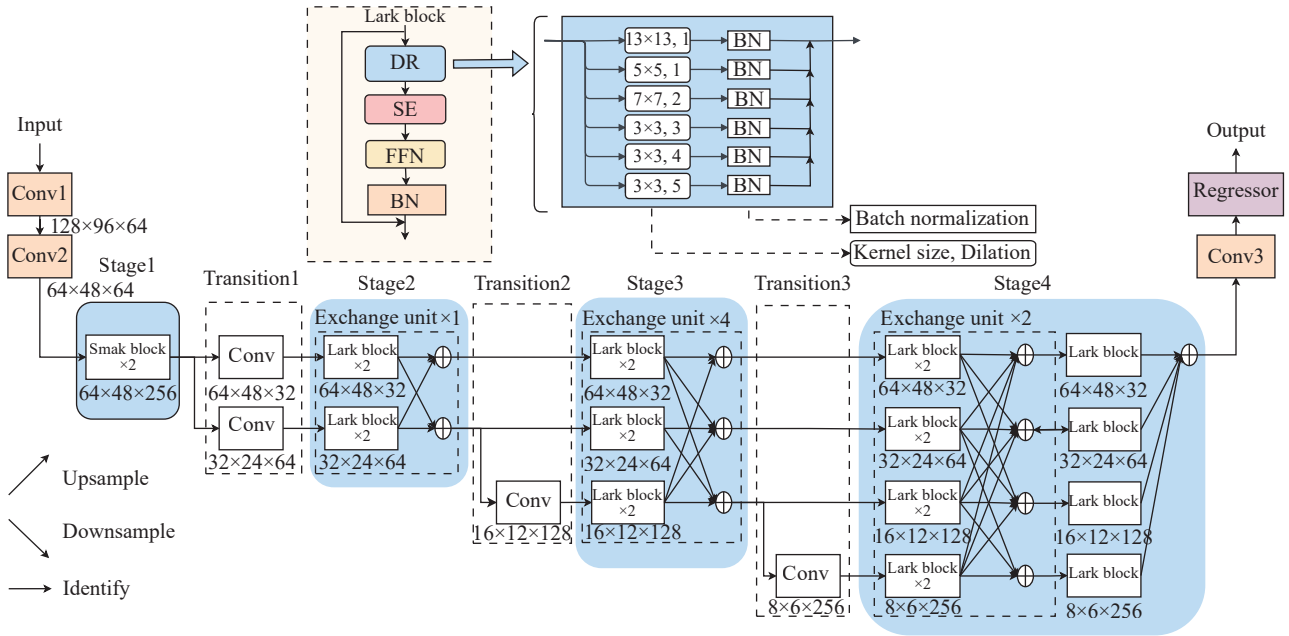


图 1 RepLK-HRNet 整体框架

Fig. 1 Framework of RepLK-HRNet

HRNet 通过并行连接不同分辨率的子网络，并在整个特征提取过程中进行多尺度特征融合，显著提高了模型的空间精确度和语义丰富度。然而，即便如 HRNet 这样的高效网络，在处理关键点遮挡、重合及复杂环境干扰等视觉任务时，仍可能面临缺少语义信息、感受野不足的问题。为此，进一步融合多层次特征提取，网络可以在不同分辨率和深度上同时学习空间信息，进一步提高空间精度。同时加入空洞重参数化卷积来优化模型。空洞卷积通过在卷积核中插入零来增加感受野，从而在不增加参数量的前提下，有效扩大卷积层的视野范围。结合重参数化技术，本文可以在训练阶段利用复杂的卷积结构来提高模型的性能，而在推理阶段则将其简化为标准的卷积层，以减少计算复杂度和模型大小。

综上所述, 本文旨在通过多尺度特征融合和多层次特征提取两个方面增强模型捕获特征的能力, 进一步提升深度学习模型在关键点遮挡、重合及复杂环境干扰等视觉任务中的性能。具体地讲, 本文将以 HRNet 为基础框架, 通过并行结构对不同分辨率 (即不同细粒度) 下的特征进行融合。同时引入大核卷积和重参数化技术, 优化特征提取过程, 对同一分辨率的输入使用不同核大小、膨胀率的空洞卷积并行提取不同抽象层次或复杂度的特征。多尺度特征涵盖了从微观到宏观的不同尺度下的特征, 小尺度特征可以捕捉细节信息, 而大尺度特征则可以反映整体结构, 二者可以相互补充, 提供更全面的信息。多层次特征应用不同参数的卷积核, 从同一分辨率下的输入数据中提取出具有不同抽象层次或复杂度的特征集合。较小的卷积核通常对局部细节敏感, 如边缘、纹理, 而较大的卷积核则能够捕捉到更广泛区域内的信息, 如形状、对象等。这些特征集合能够更全面地描述输入数据的特性, 使得模型能够在实现关键点精确定位的同时, 对被遮挡关键点有较强的推理能力。这一改进不仅提升了模型的精度和鲁棒性, 并通过重参数化技术减少参数量实现模型的轻量化, 为实际应用提供更加高效、可靠的解决方案。

2　RepLK-HRNet 模型

RepLK-HRNet 模型结构如表 1 所示, 其整体架构包括主干层(Stem)、主体和回归层(Regressor), 其中, W、H 分别表示宽度和高度, c 表示基准通道数,

表 1　RepLK-HRNet 模型结构

Table 1　Architecture configuration of RepLK-HRNet

Item	Branch 1	Branch 2	Branch 3	Branch 4
Stem	$\left[\begin{array}{c}\text{Conv}3\times3\\ \frac{W}{4}\times\frac{H}{4}\times2c\end{array}\right]\times2$			
Stage1	Downsample, $\frac{W}{4}\times\frac{H}{4}\times8c$ $\left[\begin{array}{c}\text{Smak block}\\ \frac{W}{4}\times\frac{H}{4}\times8c\end{array}\right]\times2$			
Transition1	$\left[\begin{array}{c}\text{Conv}3\times3\\ \frac{W}{4}\times\frac{H}{4}\times c\end{array}\right]$	$\left[\begin{array}{c}\text{Conv}3\times3\\ \frac{W}{8}\times\frac{H}{8}\times2c\end{array}\right]$		
Stage2	$\left[\begin{array}{c}\text{Lark block}\\ \frac{W}{4}\times\frac{H}{4}\times c\end{array}\right]\times2$	$\left[\begin{array}{c}\text{Lark block}\\ \frac{W}{8}\times\frac{H}{8}\times2c\end{array}\right]\times2$		
	Fusion			
Transition2	Identity	Identity	$\left[\begin{array}{c}\text{Conv}3\times3\\ \frac{W}{4}\times\frac{H}{4}\times4c\end{array}\right]$	
Stage3	$\left[\begin{array}{c}\text{Lark block}\\ \frac{W}{4}\times\frac{H}{4}\times c\end{array}\right]\times2$	$\left[\begin{array}{c}\text{Lark block}\\ \frac{W}{8}\times\frac{H}{8}\times2c\end{array}\right]\times2$	$\left[\begin{array}{c}\text{Lark block}\\ \frac{W}{16}\times\frac{H}{16}\times4c\end{array}\right]\times2$	
	Fusion			
Transition3	Identity	Identity	Identity	$\left[\begin{array}{c}\text{Conv}3\times3\\ \frac{W}{4}\times\frac{H}{4}\times8c\end{array}\right]$
Stage4	$\left[\begin{array}{c}\text{Lark block}\\ \frac{W}{4}\times\frac{H}{4}\times c\end{array}\right]\times2$	$\left[\begin{array}{c}\text{Lark block}\\ \frac{W}{8}\times\frac{H}{8}\times2c\end{array}\right]\times2$	$\left[\begin{array}{c}\text{Lark block}\\ \frac{W}{16}\times\frac{H}{16}\times4c\end{array}\right]\times2$	$\left[\begin{array}{c}\text{Lark block}\\ \frac{W}{32}\times\frac{H}{32}\times8c\end{array}\right]\times2$
	Fusion			
	Fusion(all up)			
Regressor	Conv1×1, p			

p 表示卷积层的输出通道数。特征提取网络以 HRNet 为基本框架，保持高分辨率输入，提取多尺度特征。同时为了给姿态估计网络提供更丰富的特征信息，考虑以大核卷积模块和空洞小核卷积并行提取多层次特征。

2.1 主干层

网络的主干层为两个核大小为 3×3，步长为 2 的卷积层，用于下采样和增加通道数，如下式所示：

$$X_0 = \mathrm{Stem}(I) \tag{1}$$

其中，$I \in R^{W\times H\times 3}$ 表示输入的图；$X_0 \in R^{\frac{W}{4}\times\frac{H}{4}\times 2c}$ 是主干层的输出。通过对图像下采样，使通道数增加到基准通道数的 2 倍，分辨率大小调整到输入图像的 1/4。

2.2 主体

主体包含 4 个特征提取阶段 (Stage) 和 3 个过渡阶段 (Transition)。Stage1 的输入分辨率大，特征提取耗时长，因此本文只使用 1 层来降低延迟。Stage2、Stage3、Stage4 分别包含 1、4、3 个特征融合单元 (Exchange unit)。整个网络共有 8 个特征融合单元，即进行了 8 次多尺度特征融合。其中每个特征融合单元包含 n(1、2、3、4)个平行分支，每个分支上包含 2 个重参数化大核卷积块和 1 个跨分辨率的融合单元。第 $m(m=1,2,\cdots,n)$个分支中特征映射的通道数和分辨率分别为第 1 个分支的 $2^{m-1}\times$ 和 $\frac{1}{2^{m-1}}\times$。

而在每个 Transition 中，引入 1 个通道数加倍 ($2^n\times$)、分辨率减半 ($\frac{1}{2^n}\times$) 的分支，为下一个 Stage 增加 1 个分支，补充不同尺度的特征信息。

2.2.1 Stage1　Stage1 采用 2 个 Smak 块，如图 2(a) 所示。Smak block 由深度可分离卷积层 (DW 层) 和 FFN 层组成，并在 DW 层和 FFN 层之间加入 SE block[34] 来增加模型深度。Smak block 可以表示为等式 (2)。

$$X_1^1 = f\left(X_1^{1(10)}\right) = X_1^{1(10)} + f'\left(X_1^{1(10)}\right) \tag{2}$$

其中，$X_1^{1(10)} = \mathrm{ds}(X_0)$，$\mathrm{ds}(\cdot)$ 表示下采样层，$X_1^{1(10)} \in R^{\frac{W}{4}\times\frac{H}{4}\times 8c}$，$X_0$ 表示通过主干层的输出；$f'(\cdot)$ 包括 3 个子模块，即 DW、SE、FFN，它们分别表示为式 (3)、(4)、(5)。

$$X_1^{1(11)} = \mathrm{BN}(\mathrm{DWconv}_3^1\left(X_1^{1(10)}\right)) \tag{3}$$

其中，$\mathrm{DWconv}_k^l(\cdot)$ 表示步长为 l 的 $k\times k$ 深度可分离卷积。卷积之后通过 SE block 和 FFN 层来增加深度，最终使输出特征图的分辨率和维度与输入特征图相同。

$$X_1^{1(12)} = \mathrm{SE}\left(X_1^{1(11)}\right) \tag{4}$$

$$X_1^{1(13)} = \mathrm{BN}(\mathrm{FFN}(X_1^{1(12)})) \tag{5}$$

其中，$X_1^{1(11)}, X_1^{1(12)}, X_1^{1(13)} \in R^{\frac{W}{4}\times\frac{H}{4}\times 8c}$

2.2.2 Transition1　Transition1 可以表示为等式 (6) 和等式 (7)。

$$Y_1^1 = \mathrm{conv}_3^1\left(X_1^1\right) \tag{6}$$

$$Y_1^2 = \mathrm{conv}_3^2\left(X_1^1\right) \tag{7}$$

其中，$X_1^1 \in R^{\frac{W}{4}\times\frac{H}{4}\times 8c}$ 表示 Stage1 最后 1 个子块的最后模块的输出特征；$\mathrm{conv}_k^l(\cdot)$ 表示步长为 l 的 $k\times k$ 卷积操作；$Y_1^1 \in R^{\frac{W}{4}\times\frac{H}{4}\times c}$，$Y_1^2 \in R^{\frac{W}{8}\times\frac{H}{8}\times 2c}$，$\{X_1^1\}$ 通过 Transition1 转换成 $\{Y_1^1, Y_1^2\}$ 两条分支。

2.2.3 其他Stages　Stage2、Stage3、Stage4 的每个阶段包含多个交换单元，这些阶段中的特征融合单元与 Stage1 有很大的不同。以 Stage3 为例，Stage3 经历了 4 个特征融合单元，每个特征融合单元中有 3 个并行分支，每个子分支通过 2 个 Lark block 提取特征，如图 2(b) 所示。

$$X_{n(e)}^m = f_{n(e)}^m(Y_{n-1}^m)\, e = 1 \tag{8}$$

$$X_{n(e)}^m = f_{n(e)}^m\left(X_{n(e-1)}^m\right) e \neq 1 \tag{9}$$

其中，函数 $f_{n(e)}^m(\cdot)$ 表示输入特征在第 n 个 Stage 对第 e 个特征融合单元的第 m 个分支所包含的一系列操作，$X_{n(e)}^m \in R^{\frac{W}{2^{m+1}}\times\frac{H}{2^{m+1}}\times 2^{m-1}c}$ 是相应的输出特征。

在子网络之后，多分辨率融合单元用于生成高分辨率特征，如式 (10) 所示：

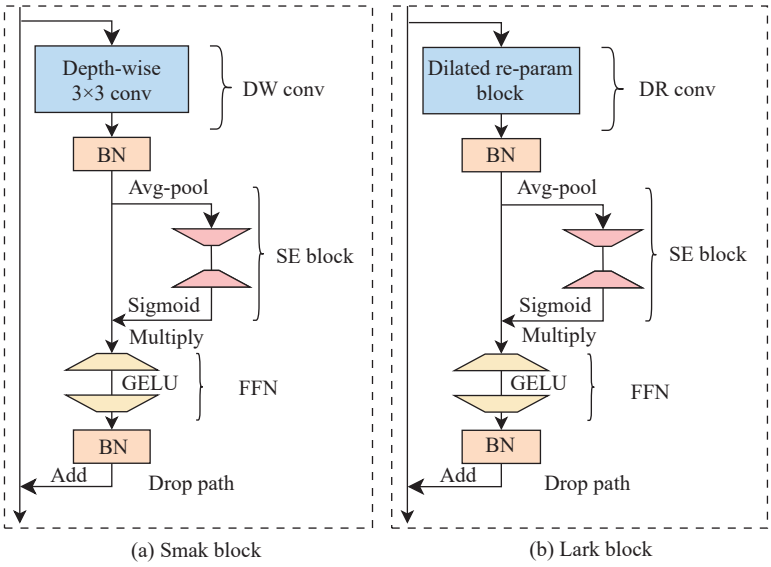
$$X_{n(e)}^m = \sigma\left(\sum_{i=1}^{m} f(X_{n(e)}^i, m)\right) \tag{10}$$

其中，$X_{n(e)}^m \in R^{\frac{W}{2^{m+1}}\times\frac{H}{2^{m+1}}\times 2^{m-1}c}$ 是第 n 阶段第 e 个特征融合单元的第 m 个多分辨率融合单元的输出，σ 是 ReLU 激活函数。如果 $i<m$，则函数 $f(X^i_{n(e)}, m)$ 表示 $2^{m-i}\times$ 下采样；如果 $i>m$，则 $f(X^i_{n(e)}, m)$ 表示 $2^{m-i}\times$ 上采样；当 $i=m$ 时，该函数表示恒等输出。

每个子网络都能提供通过多分辨率融合单元在不同分辨率的子网络之间重复收集信息的能力。第 n 个 Stage 的最后一个交换单元的输出是 $\{X_n^1, X_n^2, \cdots, X_n^m\}$。

2.2.4 其他Transitions　除了最后一个分支是由步长为 2 的 3×3 卷积组成外，其他所有分支都是恒等操作 (Identify shortcuts)。第 n 个分支转换可以表示为式 (11)、(12)。

$$Y_n^m = X_n^m,\ m = 1,2,\cdots,n \tag{11}$$



(a) Smak block　　(b) Lark block

图 2　重参数化大内核卷积块

Fig. 2　Block of reparameterized large kernel convolution

$$X_{n(e)}^{m}=\mathrm{conv}_3^2(X_n^{m-1}),\ m=n+1 \tag{12}$$

其中，$\{X_n^1,X_n^2,\cdots,X_n^m\}$ 通过 n 个 Transition 转换为 $\{Y_n^1,Y_n^2,\cdots,Y_n^m,Y_n^{m+1}\}$。Stage4 的最后一个特征融合单元仅输出 $\{X_4^1\}$。另外，基于 ResNet 中的残差块思想[1]，RepLK-HRNet 中的每个模块都引入了随机深度，即在神经网络的某些层中随机“丢弃”部分计算路径，从而减少前向传播的计算量，加速训练过程而不显著牺牲模型准确度。同时，通过增加训练时网络的变化性，增强模型的学习能力和泛化能力。

2.3　回归层

将上述网络获得的高分辨率表示输入到具有 p 个输出通道的 1×1 卷积层，得到每个关节点对应的特征图（$X\in R^{W\times H\times\{H_1,H_2,\cdots,H_p\}}$，其中 $\{H_1,H_2,\cdots,H_p\}$ 是人体关节点的置信度集合）。如在实验中数据集有 17 个关节点，则 p=17。然后使用均方误差损失函数比较真实热图和预测热图。损失函数用式 (12)、(13) 表示为：

$$\mathrm{Loss_g}=\sum_{i=1}^{4}\mathrm{MSE}(M_i,M^*) \tag{13}$$

$$\mathrm{MSE}(M_i,M^*)=\frac{1}{n\times m}\sum_{j=1}^{n}\sum_{k=1}^{m}(N_{j,k}-N_{j,k}^*)^2 \tag{14}$$

式中，$\mathrm{Loss_g}$ 为四路损失函数值之和，采用的是均方误差 (Mean Squared Error, MSE)；M_i 为第 i 个层级的关键点预测值；M^*为人工注释关键点的真实值；$N_{j,k}$、$N_{j,k}^*$分别为第 j 个人体的第 k 种关键点的预测值和人工注释关键点的真实值；n 为人体边界框个数；m 为 1 个人体的关键点个数。

3　重参数化大核卷积块

重参数化大核卷积块由重参数化大核卷积层 (DR conv) 和 FFN 组成，并在 DR 层和 FFN 层之间加入 SE block 来增加模型深度。在卷积层之后使用 BN 代替常规的 LayerNorm，BN可以等价地合并到卷积层以消除其推理成本。在 FFN 之后使用另一个 BN，也可以等效地合并到前一层 (即 FFN 中的第 2 个线性层)，如图 2(b)所示。

Stage1 使用 3×3 深度可分离卷积 (Depth-wise conv) 作为 DW 层。最后 3 个 Stage 使用 13×13 空洞重参数块 (Dilated reparam block) 作为 DR 层，空洞重参数块的具体流程如图 3 所示。

3.1　空洞重参数块

使用空洞的小内核重新参数化大内核，大内核受益于并行空洞卷积层捕获稀疏特征的能力[30]。本文使用空洞重参数块进行多层次特征提取，它使用一个非空洞的小核和多个空洞的小核层来增强一个非空洞的大核卷积层，捕获不同抽象层次的信息，增强特征表达能力。其超参数包括大核 K 的大小、并行卷积层 k 的大小以及空洞率 r。在实验中，设置 K=13，如图 3 所示，包含 5 个并行层，r=(1,2,3,4,5)，k=(5,7,3,3,3)，因此等价的核大小分别是 (5,13,7,9,11)。训练后为了减少推理成本，将整个块等价转换为一个非空洞大核卷积层，将 $r>1$ 的每个层进行适当的零填充，如式 (15) 所示。

$$w'=\mathrm{conv\_transposed}(w,I,\mathrm{stride}=r) \tag{15}$$

然后将每个 BN 合并到前面的卷积层中，将所有

得到的核相加。conv_transposed(·)表示空洞率为r、卷积核为恒等核$I \in R^{1\times1}$的转置卷积，$w \in R^{k\times k}$，$w' \in R^{((k-1)\times r+1)\times((k-1)\times r+1)}$。如图3中$k$=3, r=5的层被转换为一个稀疏的11×11核，并在每侧用一像素“0”填充，加到13×13非空洞大核上。

图4 (a)和4(b)分别展示了HRNet和RepLK-HRNet模型的有效感受野(ERF)。可以看出，RepLK-HRNet的ERF明显大于HRNet，这一结果的根本原因是增加了卷积核的大小。使用更大的卷积核使模型能够利用更大的有效感受野来捕捉到更广泛的上下文信息，学习长距离依赖关系。

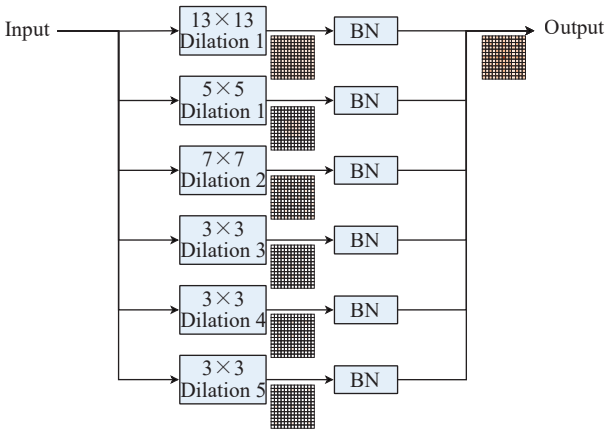
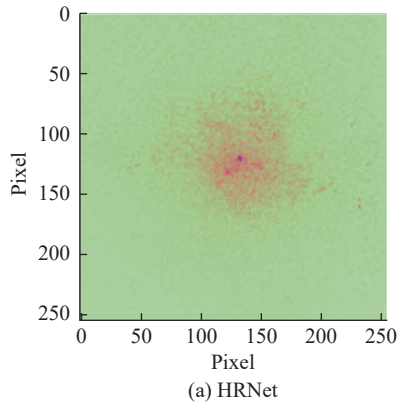
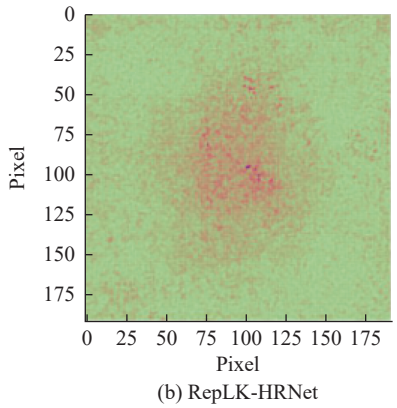


图3 空洞重参数化块的流程图

Fig. 3 Flowchart of dilated reparam block



(a) HRNet



(b) RepLK-HRNet

图4 有效感受野可视化

Fig. 4 Receptive field ERF

3.2 SE block和FFN

在增大感受野的同时，同时考虑特征抽象层次和一般表征能力。在模型设计中，浅层模型的感受野不宜过大；深层网络直接使用过大卷积核会造成浪费；使用大核卷积让模型提早获得足够感受野，不利于模型的表征能力。

因此，本文在Stage1中用3×3 Depth-wise小卷积提高特征抽象层次，如图2(b)所示。在空洞重参数块中使用一些高效结构(如SE block、FFN等)来提高模型的深度，从而增强其一般的表示能力。SE block可以在不增加网络参数和计算量的情况下，引入全局信道间的依赖关系。通过池化操作将输入特征图的空间维度压缩为一个小的特征向量，再通过激活函数和全连接层，根据特征向量学习到一个权重向量，用于对原始特征图进行加权重组合。浅层特征通常包含了更具区分性的信息，而深层特征则包含了更抽象的语义信息。SE block能够自适应地学习到每个信道的重要性，并对特征图进行动态调整。FFN的加入可以增加模型的复杂度和非线性，使其能够学习更复杂的函数映射关系，从而提高模型的表征能力和泛化能力。

4 实验结果与分析

4.1 数据集

标准人体姿态估计实验使用被广泛用于人体姿态估计的公开基准数据集MS COCO2017和MPII进行验证，被遮挡人体姿态估计实验使用OCHuman数据集进行验证。MS COCO2017分别提供了大约118 000和5 000个样本作为人体姿态估计的训练集和验证集。MPII数据集是从在线视频中提取出大约25 000张图像，每张图像包含1个或多个人，总共有超过40 000个人带有注释的身体关键点，一般将28 000个样本用作训练集，11 000个样本用作测试集。

OCHuman数据集如图5所示。该数据集聚焦于被严重遮挡的人体，提供了包括边界框、人体姿态和实例掩码在内的全面标注。该数据集包含5 081张图像中精心标注的13 360个人体实例，每个人的平均最大交并比(MaxIoU)为0.573。本文提取了其中包含关键点和掩码标注的子集(包含4 731张图片，共8 110个人)，并将其按照7∶3的比例分为训练集和验证集(训练集包含3 311张图片，共5 689个人；验证集1 420张图片，共2 421个人)。

4.2 评价指标

标准人体姿态估计实验使用由MS COCO2017

图 5　OCHuman 数据集

Fig. 5　OCHuman Datasets

数据集规定的目标关键点相似度 (Object Keypoint Similarity, OKS) 和由 MPII 数据集规定的正确关键点比例 (Percentage of Correct Keypoint, PCK) 作为模型精度的评价指标。OKS 通过计算预测关键点和其真实值的相似度来衡量, 用式 (16) 表示为:

$$\mathrm{OKS}=\frac{\sum_i \exp\left(-\frac{d_i^2}{2s^2k_i^2}\right)\delta(v_i>0)}{\sum_i \delta(v_i>0)} \tag{16}$$

式中, i 为关键点下标; d_i 为第 i 个关键点预测值与其人工注释真实值之间的欧式距离; v_i 为第 i 个关键点的可见度标志位, v_i= 1 表示第 i 个关键点无遮挡且已被注释; v_i= 2 表示第 i 个关键点有遮挡但已被注释; $\delta(\cdot)$ 为可见度判断函数; s 为关键点的衰减常数, 其值为人体边界框面积的平方根; k_i 为第 i 类关键点的归一化参考值, 是通过计算所有样本集中人工注释的关键点与其真实值之间的标准差得到, k_i 越大表示此类型的关键点越难注释。OKS 的值在 [0,1] 范围内, OKS=1 表示完美预测, 但它通常为一个范围, 当其大于或等于设定阈值 T 时, 表示预测关键点正确, 否则预测错误。

在 MPII 数据集中, 将人体头部长度作为归一化参考值, 提出 PCKh(PCK normalized by head size)评价指标, 它常用的阈值为0.5、0.2 和 0.1, 分别对应 PCKh@0.5、PCKh@0.2 和 PCKh@0.1, 预测关键点与其对应的人工注释真实值之间的归一化距离小于设定阈值, 则此关键点被视为正确预测。PCK 用式 (17) 表示为:

$$\mathrm{PCK}=\frac{\sum_i \delta\left(\frac{d_i}{d_{\mathrm{def}}}\leqslant T\right)}{\sum_i 1_i} \tag{17}$$

式中, d_{def} 为归一化参考值即人体头部长度。

在 OCHuman 数据集中, 单个人体包含 17 种关键点, 同样将 OKS 作为模型精度的评价指标。同时, 根据图像中人体被遮挡的程度将数据集划分为中等难度的实例 (0.50, 0.75) 和高等难度的实例 (0.75, 1.00), 所有实例均表示为 (0, 1.0)。针对实例中不同难度等级的数据集对模型的准确度和鲁棒性进行验证。

4.3　实验设置

具体实验环境设置如下: 本文网络结构使用深度学习框架 Pytorch 搭建。硬件配置: 操作系统为 Linux, 处理器为 7 vCPU Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40GHz, 安装内存大小为 30 G, GPU 型号为 NVIDIA GeForce RTX RTX 3080x2(20GB)。

在实验中将 RPLK-HRNet 模型的训练周期设置为 250, 批量大小设置为 32, 优化器设置为 Adam, 采用多阶梯式学习率衰减方式, 初始学习率设置为 0.001, 分别在第 125, 175, 220 个 epoch 进行学习率衰减, 衰减率设置为 0.4, 模型验证周期设置为 1 个 epoch。

4.4　实验结果与分析

4.4.1　标准人体姿态估计实验　在 MS COCO2017 数据集上进行标准人体姿态估计实验验证。以 OKS 作为模型在 MSCOCO2017 数据集上的评价指标, 通过实验统计得到在不同 OKS 阈值下的识别精度, 如表 2 所示, 其中 AP^t 为阈值为 t 时的识别精度, AP^{M} 和 AP^{L} 分别表示中等目标和大目标的平均识别精度。由表 2 可知, 本文提出的 RepLK-HRNet 模型的 OKS 相比 Stacked Hourglass 模型[5]、Simple Baselines 模型[35]、HRNet 模型[9]、HigherHRNet 模型[36]、MoveNet 模型[37] 和基于 Transformer 的 TokenPose[19]、HRFormer[21] 均有不同程度的提升, 表明 RPLK-HRNet 模型在标准人体姿态估计上表现出更好的预测性能。

表 2　MS COCO2017 数据集 OKS 对比结果

Table 2　Comparison results of OKS of MS COCO2017 data set

Model	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	Literature
Stacked Hourglass	65.5	86.8	72.3	60.6	72.6	[5]
Simple Baselines	73.7	91.9	81.1	70.3	80.0	[35]
HRNet	75.5	92.5	83.3	71.9	81.5	[9]
HigherHRNet	68.4	88.2	75.1	64.4	74.2	[36]
MoveNet	75.1	89.7	81.9	71.5	78.1	[37]
TokenPose	75.9	92.3	83.4	72.2	82.1	[19]
HRFormer	76.2	92.7	83.8	72.5	82.3	[21]
Ours	**76.9**	**95.2**	**85.0**	**74.4**	**80.5**	

图 6 所示为在 MS COCO2017 数据集上训练得到的多步学习率衰减、损失函数和精度的变化曲

线。实验采用多阶梯式学习率衰减方式，初始学习率设置为0.001，分别在第125, 175, 220个epoch进行学习率衰减，衰减率设置为0.4，4个阶段的学习率分别为[0.001 0, 0.000 4, 0.000 16, 0.000 064]。在图6(b)中可以观察到，当训练达到第125个epoch时精度曲线的上升趋势未饱和，有较大继续上升的潜力。因此，延迟学习率第1次衰减的时间将获得更好的性能。

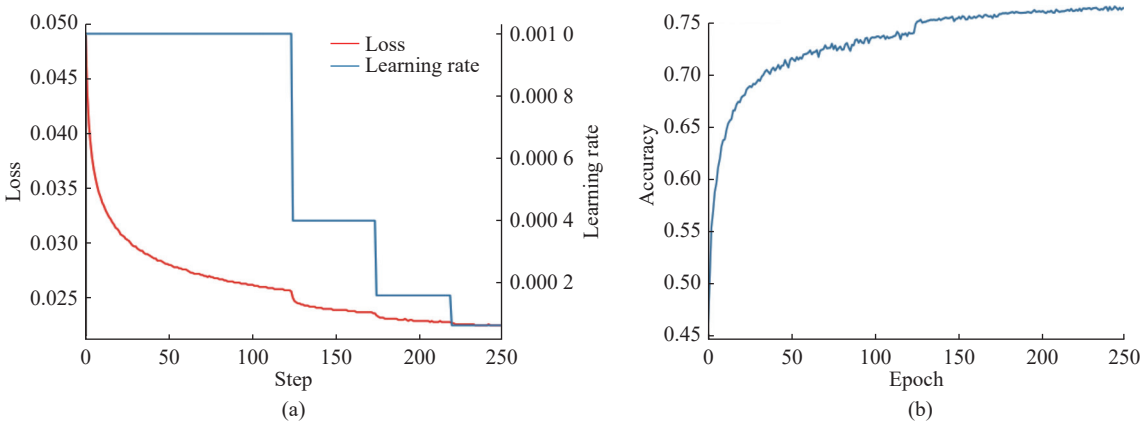


图6 训练过程中多步学习率衰减、损失函数和精度的变化曲线

Fig. 6 Change curves of multi-step learning rate decay, loss function, and accuracy during the training process

在评估本文提出的RePLK-HRNet模型在MPII数据集上的性能时，采用了PCKh@0.5作为关键指标，该指标能够全面反映模型对人体关键点检测的准确性。采用RePLK-HRNet在MPII数据集上进行标准人体姿态估计实验验证，PCKh@0.5对比实验结果如表3所示。

表3 MPII数据集PCKh@0.5对比结果

Table 3 Comparative results of PCKh@0.5 of MPII data set

Model	PCKh@0.5 of each part of the human body								Literature
	Head	Shoulder	Elbow	Wrist	Buttock	Knee	Ankle	Mean	
Stacked Hourglass	98.2	96.3	91.2	87.1	90.1	87.4	83.6	90.6	[5]
Simple Baselines	98.5	96.6	91.9	87.6	91.1	88.1	84.1	91.5	[35]
HRNet	98.6	96.9	92.8	89.0	91.5	89.0	85.7	92.3	[9]
PoseNet	98.9	72.2	93.0	98.2	86.1	63.0	44.2	79.4	[38]
GHRCPN	99.0	97.2	93.4	89.5	91.9	90.1	86.0	92.4	[39]
Ours	**99.2**	**97.5**	**93.7**	**90.0**	**92.6**	**89.5**	**87.4**	**92.6**	

由表3可知，RePLK-HRNet模型在预测各个关键点的精度上均有所提高，并且其平均预测精度也达到了较高水平。

4.4.2 *物体遮挡人体姿态估计实验* 在OCHuman数据集上进行被遮挡人体姿态估计实验验证，以OKS作为模型在OCHuman数据集上的评价指标，通过实验统计得到在不同OKS阈值下的识别精度，结果如表4所示。

对比表4结果可知，RepLK-HRNet在不同OKS阈值下都表现出较高的精度，对比HRNet，当ORS阈值为75时，RepLK-HRNet模型的精度提高更为明显。

4.4.3 *重参数化大核卷积消融实验* 本文比较了基本的HRNet模型和在特征提取网络中融入重参数化

表4 OCHuman数据集OKS对比实验结果

Table 4 Comparison results of OKS of OCHuman data set

Model	AP	AP^{50}	AP^{75}	Literature
Mask RCNN	20.2	33.2	24.5	[40]
SimpleBaseline	24.1	37.4	26.8	[41]
SPPE+	27.6	40.8	29.9	[42]
OPEC-Net	29.1	41.3	31.4	[43]
SPM	47.6	67.5	53.2	[44]
DEKR	52.2	69.9	56.6	[45]
HRNet	45.9	81.7	45.5	[9]
Ours	**56.7**	**84.6**	**59.7**	

大核卷积模型在标准和物体遮挡人体姿态估计实验中的性能, 网络的其他结构不变, 结果如表 5、表 6 所示。

HRNet 在每个特征融合单元中使用 4 层 BasicBlock[9], 而 RepLK-HRNet 在每个特征融合单元中仅用了 2 层空洞重参数块, 整个网络层数减少 $\frac{1}{2}$, 计算复杂度参数 Params 和 GFLOPs 均显著降低, 不同通道数(32,48)设置下参数量分别降低了 60.0%、63.84%。但相比于 HRNet, 改进后的模型在不同参数(256×192, 384×288, 256×192, 384×288)下的 AP 值分别提高了 0.67%, 1.60%, 1.05%, 1.83%(表 5), 可以看出, 重参数化大核卷积对不同层次的高分辨率表征有较好的提取效果。

由表 6 可知, 在标准人体姿态估计实验中, 相比 HRNet 模型, 本文模型在 MPII 数据集上的平均分数提高了 0.3。

结合表 5 和表 6 可知, RepLK-HRNet 模型比 HRNet 模型在两个典型的公开数据集上都表现出更优的关键点预测性能, 表现出较好的泛化能力。

表 7 所示为物体遮挡数据集人体姿态估计重参数化大核卷积消融实验结果。按照人体被遮挡的程度将数据集划分为完整数据集(All(0, 1.00))、中等难度数据集(Moderate(0.50, 0.75))和高等难度数据集(Hard(0.75, 1.00))。由表 7 可知, 在物体遮挡人体姿态估计实验中, 相比 HRNet 模型, RepLK-HRNet 模型在 OCHuman数据集上的不同遮挡比例下的 OKS 精度均较高, 其中, 在完整数据集中平均精度提高了 10.8。在中等难度和高等难度实例中, 平均分数分别提高了 26.9, 20.2。可以看出, 由于 RepLK-HRNet 在基础的特征提取网络中添加了空洞重参数化大核

表 5　标准数据集 MS COCO2017 人体姿态估计重参数化大核卷积消融实验

Table 5　Human pose estimation reparameterized large kernel convolutions ablation experiment on MS COCO 2017 datasets

Model	Channel number	Resolution	Params	GFLOPs	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}
HRNet	32	256×192	28.5	7.10	74.4	90.5	81.9	70.8	81.0
	32	384×288	28.5	16.0	75.1	90.6	82.2	71.5	81.8
	48	256×192	63.6	14.6	75.8	90.6	82.7	71.9	82.8
	48	384×288	63.6	32.9	76.3	90.8	82.9	72.3	83.4
RepLK-HRNet	32	256×192	**11.4**	**6.3**	**74.9**	92.5	82.5	72.3	79.1
	32	384×288	**11.4**	**14.3**	**76.3**	93.4	83.2	72.9	80.9
	48	256×192	**23.0**	**9.1**	**76.6**	93.6	83.8	73.3	80.9
	48	384×288	**23.0**	**20.5**	**77.7**	93.5	84.6	74.8	82.4

表 6　标准数据集 MPII 人体姿态估计重参数化大核卷积消融实验

Table 6　Human pose estimation reparameterized large kernel convolutions ablation experiment on MPII datasets

Model	PCKh@0.5 of each part of the human body							
	Head	Shoulder	Elbow	Wrist	Buttock	Knee	Ankle	Mean
HRNet	98.6	96.9	92.8	89.0	91.5	89.0	85.7	92.3
Ours	**99.2**	**97.5**	**93.7**	**90.0**	**92.6**	**89.5**	**87.4**	**92.6**

表 7　物体遮挡数据集 OCHuman 人体姿态估计重参数化大核卷积消融实验

Table 7　Human pose estimation reparameterized large kernel convolutions ablation experiment on OCHuman datasets

maxIouRange	RepLK-HRNet			HRNet		
	AP	AP^{50}	AP^{75}	AP	AP^{50}	AP^{75}
All(0,1.00)	56.7	84.6	59.7	45.9	81.7	45.5
Moderate(0.50,0.75)	55.0	84.7	57.9	28.1	65.3	19.3
Hard(0.75,1.00)	28.6	64.3	21.6	8.4	29.2	2.2

卷积,增加网络对远距离稀疏特征的关注能力,拥有较大的感受野,从而获得包含更有效信息的关键点特征,在实例被遮挡严重的情况下,有较好的特征提取能力,使得模型能够在实现关键点的精确定位的同时,对被遮挡关键点有较强的推理能力。

4.5 实验局限

在姿态估计任务中,除了必要的硬件配置与先进的网络算法外,庞大的数据训练库同样至关重要,它是支撑实验结果优化的基石。本文采用MS COCO2017和MPII这两个公开数据集进行标准人体姿态估计实验验证,它们提供了丰富的训练样本,有助于网络模型达到较为理想的性能表现,具有较高的泛化能力。在探索物体遮挡对人体姿态估计的影响时,采用了OCHuman数据集,尽管专注于遮挡场景,但其包含的训练与测试样本数量有限,给实验结果带来了一定的局限性。

数据集的质量对于提升模型准确性具有不可忽视的作用,但无论是通过人工标注还是借助复杂的动作捕捉设备,数据集的构建成本均相当高昂。当前,针对遮挡场景下的人体姿态估计,数据集的采集多依赖于动作捕捉技术,这种方法受限于特定环境和有限的活动范围,导致室外或更广泛场景下的遮挡人体姿态数据相对稀缺。因此,仅凭现有的公开数据集来训练模型,其性能仍有较大的提升空间,未来需要更多样化、高质量的数据集来支持这一领域的研究与发展。

5 结束语

人体姿态估计是当前计算机视觉领域的热门研究课题,现有的一些相关研究均可以有效地预测关键点,但在被遮挡的场景下却难以达到准确识别关键点的目的。本文提出了一个结合重参数化大核卷积的高分辨率人体姿态估计模型RepLK-HRNet用于解决物体遮挡人体姿态估计问题,将重参数化大核卷积块融入特征提取网络HRNet中,使模型能够同时关注多尺度、多层次的特征,从而提取到更加丰富的信息来正确定位被遮挡关键点。实验结果表明,RepLK-HRNet模型在标准数据集和关节点被遮挡数据集上的人体姿态估计结果均有较高的准确性和较强的鲁棒性。并且,通过调整优化网络结构,模型参数量和计算复杂度都得到了显著降低,实现了模型的轻量化,有利于后续模型的边缘化部署能力。未来的研究将聚焦于获取包含更多遮挡人体样本的数据集来进一步提升模型的精度,并探索模型在边缘设备上的高效部署策略。

参考文献:

[1] HE K, ZHANG X, REN S, *et al*. Deep residual learning for image recognition[C]//2016 IEEE Conferenceon Computer Vision and Pattern Recognition(CVPR). Seattle, WA: IEEE, 2016: 770-778.

[2] XIE S, GIRSHICK R, DOLLAR P, *et al*. Aggregated residual transformations for deep neural networks[C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI: IEEE, 2017: 5987-5995.

[3] GAO S H, CHENG M M, ZHAO K, *et al*. Res2Net: A new multi-scale backbone architecture[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43: 652-662.

[4] HU J, SHEN L, SUN G, *et al*. Squeeze-and-excitation networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 42(8): 2011-2023.

[5] NEWELL A, YANG K, DENG J, *et al*. Stacked hourglass networks for human pose estimation[C]//European Conference on Computer Vision (ECCV). Switzerland: Springer Cham, 2016: 483-499.

[6] CAI Y, WANG Z, LUO Z *et al*. Learning delicate local representations for multi-person pose estimation[C]// European Conference on Computer Vision(ECCV). Switzerland: Springer Cham, 2020: 455-472.

[7] CHEN Y, WANG Z, PENG Y, *et al*. Cascaded pyramid network for multi-person pose estimation[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, UT: IEEE, 2018: 7103-7112.

[8] LIU W, CHEN J, LI C, *et al*. A cascaded inception of inception network with attention modulated feature fusion for human pose estimation[C]//The Thirty-Second AAAI Conference on Artificial Intelligence(AAAI). New Orleans, Louisiana: AAAI, 2018: 7170-7177.

[9] SUN K, XIAO B, LIU D, *et al*. Deep high-resolution representation learning for human pose estimation[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA: IEEE, 2019: 5686-5696.

[10] VASWANI A, SHAZEER N, PARMAR N, *et al*. Attention is all you need[EB/OL]. (2017-06-12) [2024-06-05]. https://arxiv.org/abs/1706.03762v2.

[11] WANG W H, XIE E Z, LI X, *et al*. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions[C]//IEEE International Conference on Computer Vision (ICCV). Montreal, QC: IEEE, 2021: 2380-7504.

[12] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, *et al*. An image is worth 16×16 words: Transformers for image

recognition at scale[EB/OL]. (2020-10-22) [2024-06-05]. https://arxiv.org/abs/2010.11929v2.

[13] CHEN H, WANG Y, GUO T Y, *et al*. Pre-trained image processing transformer[C]//IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Nashville, TN: IEEE, 2021: 12299-12310.

[14] LIANG J Y, CAO J Z, SUN G L, *et al*. SwinIR: Image restoration using swin transformer[C]//IEEE International Conference on Computer Vision Workshops (ICCVW). Montreal, BC, Canada: IEEE, 2021: 1833-1844.

[15] WANG W H, XIE E Z, LI X, *et al*. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions[C]//IEEE International Conference on Computer Vision (ICCV). Montreal, QC, Canada: IEEE, 2021: 2380-7504.

[16] XIE E Z, WANG W H, YU Z D, *et al*. SegFormer: Simple and Efficient design for semantic segmentation with transformers[EB/OL]. (2021-05-31) [2024-06-05]. https://arxiv.org/abs/2105.15203.

[17] DAI X Y, CHEN Y P, XIAO B, *et al*. Dynamic head: Unifying object detection heads with attentions[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, TN, USA: IEEE, 2021: 7369-7378.

[18] LIU Z, LIN Y T, CAO Y, *et al*. Swin transformer: Hierarchical vision transformer using shifted windows[C]//IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, QC, Canada: IEEE, 2021: 9992-10002.

[19] LI Y, ZHANG S K, WANG Z C, *et al*. TokenPose: Learning keypoint tokens for human pose estimation[C]//IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, QC, Canada: IEEE, 2021: 11293-11302.

[20] YANG S, QUAN Z B, NIE M, *et al*. TransPose: Keypoint localization via transformer[C]//IEEE International Conference on Computer Vision(ICCV). Montreal, QC, Canada: IEEE, 2021: 11782–11792.

[21] YUAN Y H, FU R, HUANG L, *et al*. HRFormer: High-resolution transformer for dense prediction[EB/OL]. (2021-10-18)[2024-06-05]. https://arxiv.org/abs/2110.09408.

[22] XU Y F, ZHANG J, ZHANG Q M, *et al*. ViTPose: Simple vision transformer baselines for human pose estimation[EB/OL]. (2022-04-26)[2024-06-05]. https://arxiv.org/abs/2204.12484.

[23] HAN Q, FAN Z J, DAI Q, *et al*. Demystifying local vision transformer: Sparse connectivity, weight sharing, and dynamic weight[EB/OL]. (2021-06-08) [2024-06-05]. https://arxiv.org/abs/2106.04263v2.

[24] ROMERO D W, KUZINA A, BEKKERS E, *et al*. Ckconv: Continuous kernel convolution for sequential data[EB/OL]. (2021-02-04)[2024-06-05].https://arxiv.org/abs/2102.02611.

[25] ROMERO D W, BRUINTJES R J, TOMCZAK J M, *et al*. Flexconv: Continuous Kernel convolutions with differentiable kernel sizes[EB/OL]. (2021-10-15) [2024-06-05]. https://arxiv.org/abs/2110.08059.

[26] SZEGEDY C, IOFFE S, VANHOUCKE V, *et al*. Inception-v4, inception-resnet and the impact of residual connections on learning[EB/OL]. (2016-02-23)[2024-06-05]. https://arxiv.org/abs/1602.07261.

[27] SZEGEDY C, LIU W, JIA Y Q, *et al*. Going deeper with convolutions[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA: IEEE, 2015: 1-9.

[28] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2014-09-04)[2024-06-05]. https://arxiv.org/abs/1409.1556.

[29] DING X, ZHANG X Y, MA N N, *et al*. Repvgg: Making vgg-style convnets great again[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Nashville, TN, USA: IEEE, 2021: 13728-13737.

[30] DING X, ZHANG X, ZHOU Y, *et al*. Scaling up your Kernels to 31×31: Revisiting large kernel design in CNNs[EB/OL]. (2022-03-13)[2024-06-05]. https://arxiv.org/abs/2203.06717.

[31] DING X, ZHANG Y, GE Y, *et al*. UniRepLKNet: A universal perception large-kernel ConvNet for audio, video, point cloud, time-series and image recognition [EB/OL]. (2023-11-27)[2024-6-5]. https://arxiv.org/abs/2311.15599.

[32] HU M, FENG J Y, HUA J S, *et al*. Online convolutional re-parameterization[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, LA, USA: IEEE, 2022: 558-567.

[33] CAI Z C, DING X H, SHEN Q , *et al*. RefConv: Re-parameterized refocusing convolution for powerful ConvNets[EB/OL]. (2023-10-16)[2024-06-05]. https://arxiv.org/abs/2310.10563.

[34] HU J, SHEN L, ALBANIE S, *et al*. Squeeze-and-excitation networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 42(8): 2011-2023.

[35] XIAO B, WU H, WEI Y C. Simple baselines for human pose estimation and tracking[C]//European Conference on Computer Vision (ECCV). Switzerland: Springer Cham, 2018: 472-487.

[36] CHENG B, XIAO B, WANG J, *et al*. HigherHRNet: Scale-aware representation learning for bottom-up human pose estimation[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, 2020: 5386–5395.

[37] JO B J, KIM S K. Comparative analysis of OpenPose, PoseNet, and MoveNet models for pose estimation in mobile devices[J]. Traitement du Signal, 2022, 39(1): 119-124.

[38] KOSEI Y, RYOSUKE K. Development of human pose recognition system by using raspberry PI and posenet model[C]//20th International Symposium on Communications and Information Technologies (ISCIT). Tottori, Japan: ISCIT, 2021: 41-44.

[39] 罗梦诗, 徐杨, 叶星鑫. 基于轻量型高分辨率网络的被遮挡人体姿态估计 [J]. 武汉大学学报 (理学版), 2021, 67(5): 403-410.

[40] HE K, GKIOXARI G, GIRSHICK R, *et al*. Mask R-CNN[C]//IEEE International Conference on Computer Vision(ICCV). Venice, Italy: IEEE, 2017: 2961-2969.

[41] XIAO B, WU H P, WEI Y C, *et al*. Simple baselines for human pose estimation and tracking[C]//European Conference on Computer Vision (ECCV). Switzerland: Springer Cham, 2018: 472-487.

[42] LI J F, WANG C, ZHU H, *et al*. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Long Beach, CA, USA: IEEE, 2019: 10863-10872.

[43] QIU L T, ZHANG X Y, LI Y R, *et al*. Peeking into occluded joints: A novel framework for crowd pose estimation[C]//European Conference on Computer Vision(ECCV). Scotland, GLASGOW : Springer Cham, 2020: 488-504.

[44] NIE X C, FENG J S, ZHANG J F, *et al*. Single-stage multi-person pose machines[C]//In Proceedings of the IEEE International Conference on Computer Vision(ICCV). Seoul, Korea (South): IEEE, 2019: 6951-6960.

[45] GENG Z G, SUN K, XIAO B, *et al*. Bottom-up human pose estimation via disentangled keypoint regression[EB/OL]. (2021-04-06)[2024-06-05].https://arxiv.org/abs/2104.02300.

High-Resolution Pose Estimation Based on Reparameterized Large Kernel Convolution

CHEN Jiayi, HUANG Xiaoyu, WU Shengxi, WANG Xuewu

(Key Laboratory of Smart Manufacturing in Energy Chemical Process, Ministry of Education, East China University of Science and Technology, Shanghai 200237, China)

Abstract: Although significant progress has been made in the field of human pose estimation, it still faces enormous challenges for achieving high-precision and robust pose estimation for the case of dynamic scene changes, occlusions, and complex backgrounds. To address these issues—particularly keypoint occlusion, overlap, and interference from complex environments—this paper proposes a high-resolution human pose estimation model incorporating large kernel convolution techniques, named RepLK-HRNet. The core innovation of the proposed model lies in its unique design of the feature extraction network, which introduces a reparameterized large kernel convolution strategy to enhance the model's ability in capturing multi-scale and multi-level feature information. Meanwhile, the network architecture is optimized to significantly reduce the number of parameters and computational complexity. Experimental results demonstrate that, compared to the traditional HRNet model, the RepLK-HRNet model achieves an improvement of 1.83% in accuracy on the standard MS COCO 2017 dataset and an increase of 23.7% in accuracy on the occlusion dataset OCHuman, while reducing Params by 63.84% and GFLOPs by 37.69%. These results indicate that RepLK-HRNet significantly improves pose estimation accuracy under general, occluded, and keypoint-confused conditions, showcasing excellent robustness and generalization capabilities. Moreover, it meets practical application demands in terms of computational efficiency and memory usage.

Key words: pose estimation; reparameterized large kernel convolution; HRNet; receptive field; feature fusion

（责任编辑：张欣）