文章编号：1006-3080(2025)04-0514-08 DOI：10.14135/j.cnki.1006-3080.20240731002

基于高阶空间交互作用的姿态估计网络

黄晓宇， 陈佳艺， 吴艺玮， 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：人体姿态估计是计算机视觉领域的一个重要研究方向。随着深度学习技术的进步，现有的姿态估计模型在预测人体关键点方面已经取得了显著成效，然而，在处理复杂场景如严重遮挡、复杂背景、极端姿态、多尺度变化和光照变化时，这些模型仍然面临挑战，准确度往往受到影响。为解决这个问题，本文提出了一种改进的基于高分辨率网络(High-Resolution Network，HRNet)的人体姿态估计方法，该方法通过引入高阶空间交互和注意力机制，显著提升了模型在复杂场景中的表现；并采用递归门控卷积和卷积注意力模块以增强模型在高阶空间特征提取的能力。结果表明，提出的方法在COCO2017数据集上超越了现有主流方法，实现了更高的姿态估计精度。

关键词：姿态估计；高分辨率网络；高阶空间交互；CBAM注意力机制；特征提取

中图分类号：TP273 **文献标志码：**A

人体姿态估计作为计算机视觉领域的一个重要研究方向，致力于从图像或视频数据中精确检测和识别人体的关键点(如头部、肩膀、肘部、手腕、髋部、膝盖和脚踝)，并推断出整体的人体姿态[1]。这项技术在虚拟现实、人机交互[2]、行为分析[3]、医疗康复和视频监控[4]等多个领域中具有广泛应用。目前的姿态估计方法主要依赖于深度卷积神经网络(CNN)[5]，这些网络在捕捉局部特征和短程依赖性方面表现优异。然而，人体姿态估计任务常常涉及复杂的空间关系和长程依赖性，例如不同关节之间的相互关系以及整体姿态的一致性，特别是在处理人体遮挡和复杂背景时。在这种背景下，高阶空间交互显得尤为重要，通过捕捉图像中的复杂和高级的空间关系，可以显著提升姿态估计的精度和鲁棒性。

近年来，深度学习，尤其是卷积神经网络的应用，显著提升了姿态估计的精度和效率[5-6]。例如，Hourglass网络首次引入了高分辨率特征恢复，通过将从高到低和从低到高的卷积块串联为基本模块，进而实现高分辨率的恢复[7]。U-Net[8]、DeconvNet[9]和ConvSegNet[10]等网络也采用了从低分辨率恢复到高分辨率的输出分类方法。SimpleBaseline[11]则利用转置卷积层生成高分辨率表示。为了处理遮挡关键点、不可见关键点和拥挤背景问题，MSPENet[12]引入了多尺度融合机制，通过将不同尺度的特征合并来进行网络训练。这些方法增强了姿态估计网络的性能，但它们主要关注不同尺度特征的融合，而未充分考虑特征融合过程中来自其他层的大量不相关信息的潜在集成。高分辨率网络(High-Resolution Network，HRNet)提出了一种创新的架构，保持整个过程中以高分辨率表示，并通过在并行多分辨率子网之间交换信息，实现了连续的多尺度融合[13]。然而，传统HRNet结构在处理复杂空间交互时可能存在一定局限性，例如对复杂空间交互的建模不足以及全局一致性处理的不足。

在特征提取网络中引入注意力机制显著提升了姿态估计任务的性能[14-15]。通道空间注意力模块(Convolutional Block Attention Module，CBAM)结合了通道注意力和空间注意力，通过提升卷积神经网

收稿日期：2024-07-31
基金项目：国家自然科学基金 (62076095)
作者简介：黄晓宇 (2000—)，男，河南鹤壁人，硕士生，研究方向为机器学习与人工智能。E-mail：1115489775@qq.com
通信联系人：吴胜昔，E-mail：wushengxi@ecust.edu.cn
引用本文：黄晓宇，陈佳艺，吴艺玮，等. 基于高阶空间交互作用的姿态估计网络 [J]. 华东理工大学学报(自然科学版)，2025，51(4)：514-521.
Citation： HUANG Xiaoyu, CHEN Jiayi, WU Yiwei, *et al*. Pose Estimation Network Based on High-Order Spatial Interactions[J]. Journal of East China University of Science and Technology, 2025, 51(4): 514-521.

络的特征表示能力, 改善了模型在视觉任务中的表现[16]。SENet(Squeeze-and-excitation networks)引入了 Squeeze-and-Excitation 机制, 以自适应调整通道特征响应, 从而提高深度神经网络的表示能力和准确性[16]。RANet(Region attention network)则专注于目标检测和分割任务, 通过区域注意力机制有效提升对图像中重要区域的关注度, 从而改善模型的性能和泛化能力[17]。这些注意力机制的引入能够帮助网络更精确地捕捉人体姿态中的细微特征变化, 进而提高姿态估计的精度和鲁棒性, 同时在复杂环境和动态场景中展现出更好的稳定性和可靠性。

综上所述, 本文以 HRNet 作为基本网络框架, 结合递归门控卷积(GnBlock)[18] 与深度可分离卷积[19], 提出了一种改进的模块 GnBlock, 并在残差模块的瓶颈层之后引入了 CBAM 注意力机制。此改进旨在进一步优化特征提取过程, 重点处理姿态关键点之间的重要关联性和空间依赖关系。

1　相关工作

1.1　高分辨率网络 HRNet

HRNet 是一种专为人体姿态估计及其他视觉任务设计的先进神经网络架构, 由 Sun 等[14] 在 2019 年提出。该架构旨在以最小的计算开销实现高分辨率输出, 从而优化其实用性, 它显著提升了人体姿态估计领域的技术水平。HRNet 在捕捉高保真特征和整合语义信息方面表现卓越, 极大地提高了模型的准确性和鲁棒性, 已成为计算机视觉和人机交互研究中的关键工具。

HRNet 的核心理念是多分辨率表示学习。它从高分辨率图像开始, 通过下采样生成多分辨率特征图。这些特征图在网络的不同阶段通过并行卷积独立处理, 以确保高分辨率数据的保留。最终, 这些特征图被融合, 以生成高分辨率的结果。这种设计不仅保证了计算效率, 还能够精确捕捉输入图像的细节, 在姿态估计任务中表现出色。HRNet网络结构如图 1 所示。

1.2　递归门控卷积 GnConv

传统的特征提取方法主要依赖于卷积操作。在 HRNet 网络模型中, 基础组件是大量的普通卷积层。然而, 卷积层的过度使用会导致计算量和参数量的显著增加。此外, 常规的卷积结构往往未能充分考虑到相邻空间区域之间的交互特性, 导致普通卷积在捕捉特征空间之间的相互作用方面存在一定的局限性。

为了解决这个问题, 本文引入了 GnConv。GnConv 由 Rao 等[20] 在 2022 年提出。该模块借鉴了 Transformer模型中的自注意力(Self-attention)机制[21], 通过结合门控卷积和递归设计, 实现了高阶空间交互。

与传统的卷积结构不同, GnConv 在特征提取过程中通过邻接空间区域的相乘操作来增强特征的表示能力。这种设计使得两个或多个普通卷积结构可以进行高阶空间交互, 从而在特征表达上达到更高的精度。具体而言, GnConv 通过门控机制动态调整特征图的空间响应, 进而在特征的相互作用中融入更丰富的信息。这种高阶的空间交互能力显著提升了卷积网络在复杂视觉任务中的表现, 使得模型能够更有效地捕捉细微的空间关系。同时, 递归设计使得模块能够有效地处理多尺度信息, 优化了特征的表达能力, 而不引入大量额外的计算开销。普通卷积结构、Transformer 模型中引入 Self-attention 机

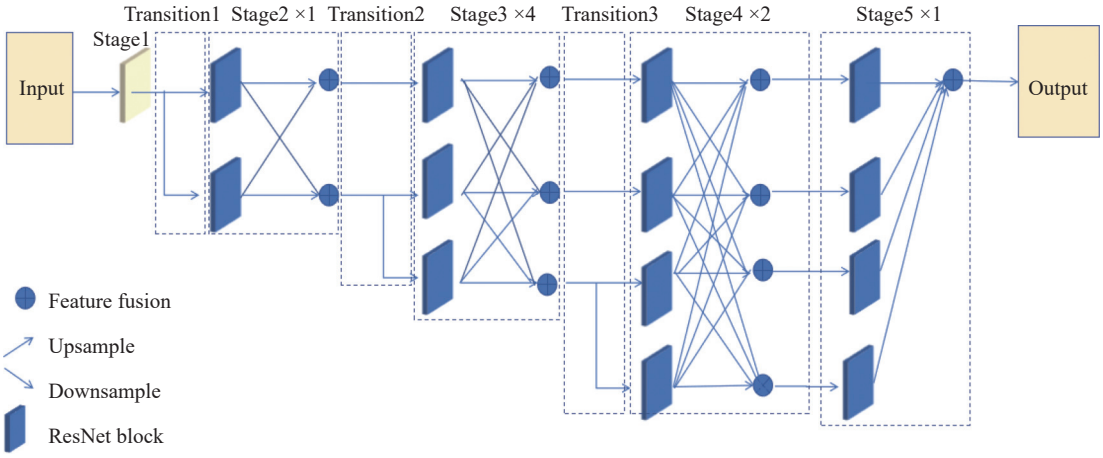
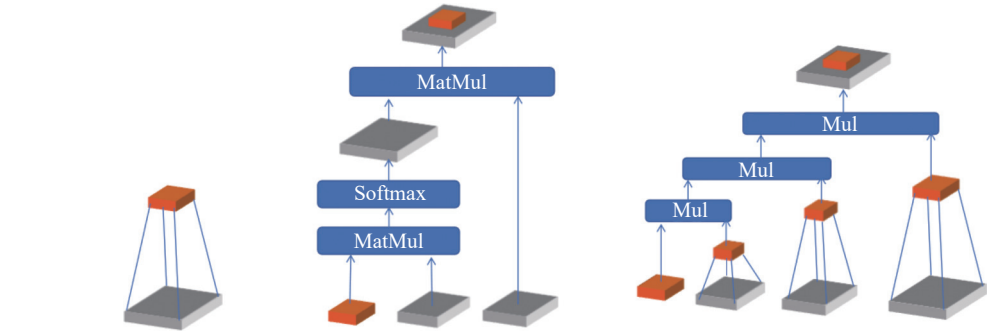


图 1　HRNet 网络结构图

Fig. 1　Framework of HRNet

制的 Transformer-block 模块[22-25]和 GnConv 的对比图如图 2 所示，其中图 2(b)示出了引入 Self-attention 机制的 Transformer-block 模块，图 2(c)为 GnConv 示意图。

普通卷积单元没有空间的交互特性，Self-attention 机制只有两个连续并且相邻的特征矩阵向量才具有交互特性，而对于 GnConv 来讲，两个相邻卷积单元以及多个卷积单元之间可以通过矩阵相乘的方式实现高阶交互。关于 GnConv 的公式见式(1)~式(5)。



(a) Standard convolution operation (b) Self-attention operation (c) Gated convolutions and recursive deigns

图 2 卷积单元对比图

Fig. 2 Comparison of convolutional units

当输入特征为 $x \in R^{H\times W\times C}$ 时，获得投影特征 p_0 和 q_0:

$$[p_0^{H\times W\times C_0}, q_0^{H\times W\times C_0}, \cdots, q_{n-1}^{H\times W\times C_{n-1}}] = \phi_{\text{in}}(x) \in R^{H\times W\times (C_0+\sum_{0\leqslant k\leqslant (n-1)} C_k)} \tag{1}$$

其中，C 代表输出通道数。

ϕ_{in} 为通道混合时的输入投影层，当执行一阶交互时：

$$p_0 = f(q_0)\odot p_0 \in R^{H\times W\times C} \tag{2}$$

其中，⊙表示点积运算；f 为深度可分离卷积 DWConv 运算，之后进行多阶交互运算：

$$p_{k+1} = f_k(q_k)\odot \frac{g_k(f_k)}{\alpha}, k = 0,1,\cdots,n-1 \tag{3}$$

其中，α 表示经过卷积操作后输出缩放的比例。将输出缩放为 $1/\alpha$ 稳定训练，然后根据不同顺序来匹配通道维度:

$$g_k = \begin{cases} \text{Identity}, & k = 0 \\ \text{Linear}(C_{k-1}, C_k), & 1 \leqslant k \leqslant n-1 \end{cases} \tag{4}$$

然后，将递归之后的 q_n 输入给通道混合时的输出投影层 ϕ_{out}：

$$y = \phi_{\text{out}}(p_{k+1}) \in R^{H\times W\times C} \tag{5}$$

GnConv 的输入是具有 C 个通道的特征图，在第 1 层卷积后，通道数翻倍。第 1 层卷积的输出被分成两部分：第 1 部分由下一层使用，第 2 部分经过深度可分离卷积，输出 3 部分作为其他 3 层的输入。这种设计增强了特征表征能力，而不引入额外的计算复杂性，其整体结构如图 3 所示。

在 HRNet 中引入递归门控卷积，可以有效地捕捉图像数据中的上下文关系和高阶特征交互。这种结合可以提升模型对复杂图像特征的理解和表征能力，从而改善任务的性能。

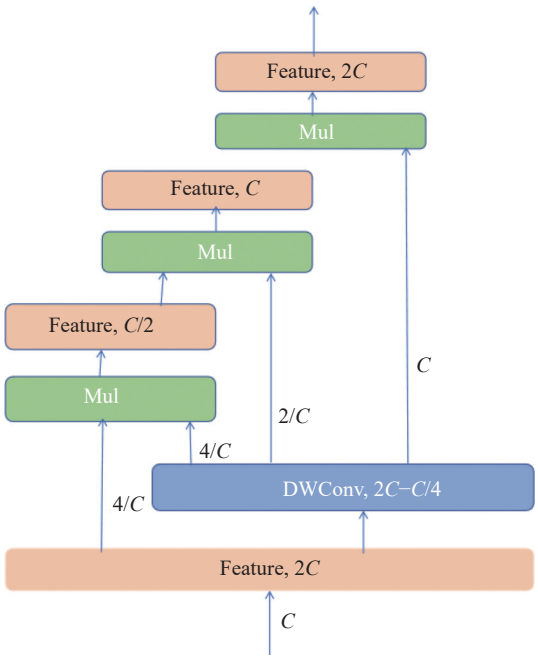


图 3 递归门控卷积结构图

Fig. 3 Structure of recursively gated convolutional

1.3 通道空间注意力模块 CBAM

为了更加有效地提高对图像中关键区域的关注度，从而改善模型的性能和泛化能力，我们引入了 CBAM 注意力机制。这个机制使得网络能够更精确地捕捉人体姿态中的细微特征变化，提高姿态估计

的准确性和鲁棒性, 在复杂环境和动态场景中展现出更加稳定和可靠的表现。

CBAM 主要由通道注意力模块(Channel Attention Module, CAM)和空间注意力模块(Spatial Attention Module, SAM)组成。CAM 模块通过全局平均池化和全局最大池化生成特征图的通道特征向量。随后, 通过多层感知机(MLP)对这些特征向量进行加权组合, 以计算每个通道的注意力权重。CAM 的核心目标是捕捉特征图中通道之间的依赖关系, 调整各通道的相对重要性。SAM 专注于捕捉特征图内不同空间位置之间的关系。它通过对通道维度进行汇聚, 生成空间特征图, 并通过卷积操作计算空间注意力分数。SAM 旨在突出图像中具有重要意义的空间区域, 抑制不相关的背景信息。

对于输入特征图 $X \in R^{C\times H\times L\times W}$, 首先经过 CAM, 得到 $X' \in R^{C\times H\times L\times W}$:

$$\boldsymbol{M}_{\mathrm{C}}(X)=\sigma(\mathrm{MLP}(\mathrm{Maxpool}(X)+\mathrm{MLP}(\mathrm{Avgpool}))) \quad (6)$$

$$X'=\boldsymbol{M}_{\mathrm{C}}(X)\otimes X \quad (7)$$

其中, $\boldsymbol{M}_{\mathrm{C}}(X)\in R^{C\times 1\times 1\times 1}$, MLP 表示多层感知机, Maxpool 表示空间域的全局最大池化, Avgpool 表示空间域的全局平均池化, σ 为 Sigmoid 函数, $\otimes$ 表示逐元素乘法。将输出的特征权重向量 $\boldsymbol{M}_{\mathrm{C}}(X)$ 重新加权到初始特征图 X 上得到 X',完成特征图在通道维度上的重标定。然后, 将 X'输入到空间注意力模块得到 $X'' \in R^{C\times H\times L\times W}$:

$$\boldsymbol{M}_{\mathrm{S}}(X')\sigma(\mathrm{conv}_{7\times 7\times 7}([\mathrm{Maxpool}(X');\mathrm{Avgpool}(X')])) \quad (8)$$

$$\boldsymbol{X''}=\boldsymbol{M}_{\mathrm{S}}(X')\otimes X' \quad (9)$$

其中, $\boldsymbol{M}_{\mathrm{S}}(X')\in R^{1\times H\times L\times W}$, $\mathrm{conv}_{7\times 7\times 7}$ 表示核尺寸为 7×7×7 的卷积运算, Maxpool 表示通道域的全局最大池化, Avgpool 表示通道域的全局平均池化, 其结构如图 4、图 5 所示。

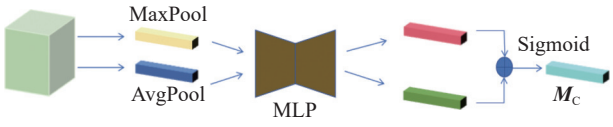


图 4　通道注意力机制结构

Fig. 4　Structure of CAM

尽管 GnConv 能够有效引入高阶空间交互操作, 提升特征提取的能力, 但它在处理通道层面依赖关系方面存在一定的不足。CAM 则能够弥补这个缺陷, 它专注于学习通道间的依赖关系, 并根据这些关系调整通道特征, 从而在通道层面优化特征提取过程。

结合 GnConv 与 CBAM, 可以充分发挥两者的

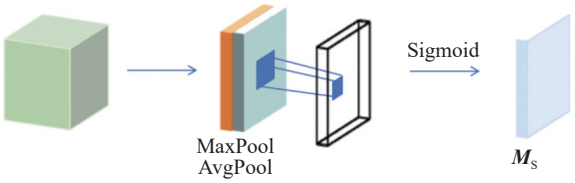


图 5　空间注意力机制结构

Fig. 5　Structure of SAM

优势。GnConv 引入的高阶空间交互操作增强了特征图在空间维度的表达能力, 而 CBAM 则在通道维度和空间维度上优化了特征的表达。两者的结合不仅提升了特征提取的全面性, 还确保了模型在捕捉复杂视觉信息时的准确性和鲁棒性。通过引入 CBAM 注意力模块, 可以有效弥补 GnConv 在通道层面上的不足, 实现更全面的特征优化, 从而提升模型在各种视觉任务中的表现。

1.4　高阶空间交互姿态估计网络

本文重新设计了 HRNet 中的 BottleNeck Block 结构, 并引入了 GnConv, 用以减少参数量并增强特征提取的高阶空间交互能力。然而, GnConv 在通道层面上的依赖关系处理方面存在一定的不足。

为弥补这个不足, 本文进一步引入了 CBAM 注意力机制, CBAM 通过学习通道间的依赖关系并对通道特征进行调整, 在通道层面优化特征提取效果。改进的姿态估计网络结构如图 6 所示。

模型包括 4 个阶段, 从高分辨率卷积流 $W\times H\times C$ 开始, 第 1 阶段包括提取图像特征和下采样, 然后进行 3×3 卷积生成额外的分辨率路径 $W/2\times H/2\times 2C$。两个不同的分辨率路径并行输出到下一个阶段。第 2、3、4 阶段分别由 1、4、3 个阶段模块组成。每个 Stage Module 包含多个不同的分辨率路径, 实现来自不同分辨率的特征的融合。Stage Module 由基本块、注意力模块 CBAM 和融合单元组成。不同分辨率的特征首先经过 4 个 Basic Block, 然后经过 CBAM 模块。最后, 通过融合单元与其他分辨率的特征进行融合。

Basic Block 由 4 个残差单元组成, 每个残差单元包含 2 个 3 × 3 卷积, 后接 BN 和 GeLU[26]。在两个阶段之间使用过渡模块来添加分辨率减半的额外路径。在 CBAM 模块中, 先应用全局平均池化和最大池化生成通道特征向量, 通过多层感知机计算通道注意力权重, 对经过通道注意力调整的特征图应用空间注意力机制, 最终得到优化后的特征图。融合单元以全连接的方式连接不同分辨率的输出, 如图 6 所示。其中, GnBlock 是基于 GnConv 操作构建的一个新的模块。GnBlock 受广泛应用于 Transformers

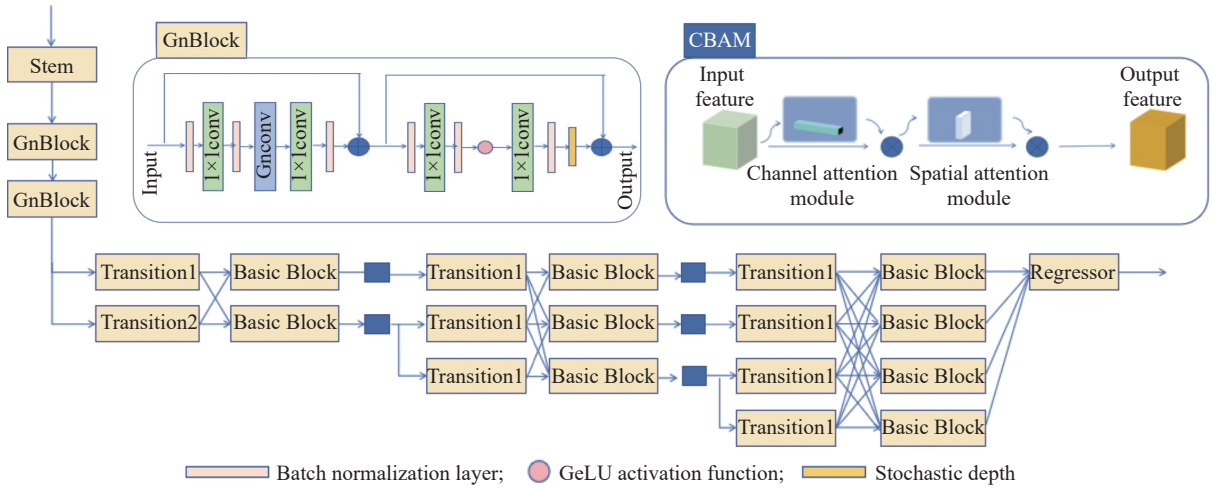


图 6 改进的姿态估计网络结构图

Fig. 6 Structure of improved pose estimation network

和 MLP 的前馈网络(FFN)的启发,结合了 CNN 的结构特性。该模块包括短路连接(Shortcut)、批量归一化(BN)层、两个 1 × 1 卷积层以及 GeLU 激活函数。相较于经典 FFN 中使用全连接层之前的层归一化(Layer normalization),BN 的优势在于能够直接集成到卷积操作中,从而提升推断效率,并优化性能。GnBlock 结构图见图 7。

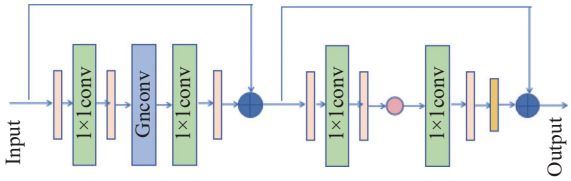


图 7 GnBlock 结构图

Fig. 7 Structure of GnBlock

2 结果和分析

2.1 数据集与评价指标

本文根据 HRNet 模型选取大型数据集 COCO2017 人体姿态估计数据集,COCO2017 包含 200 000 张图像,涵盖了 250 000 个人体样本。每个人体实例都按照指定的顺序标注了 17 个关键点,包括 5 个面部标志和 12 个身体关节。每个关键点都具有 x、y 坐标和可见性标志。本文模型在 COCO2017 验证集和测试集上进行测试评估,评价标准使用 COCO 官方提供的 OKS(Object keypoint similarity),如式(10)所示。

$$\mathrm{OKS}_P=\frac{\sum_i \exp\left(\frac{-d_{p_i}^2}{2S_p^2\sigma_i^2}\right)\delta(V_{p_i}>0)}{\sum_i \delta(V_{p_i}>0)} \tag{10}$$

其中,P 代表图像中某人 ID;P_i 代表某人关键点 ID;d_{p_i} 代表预测值与真实值之间的欧氏距离;S_P^2 代表某个人所占面积;σ_i^2 表示归一化因子;δ 表示关键点函数,且输出只有 0 或 1。

COCO 数据集一般包含平均准确度(Average Precision, AP)和平均召回率(Average Recall, AR)两个重要指标,可通过计算预测值关键点与真值关键点的相似性得到 AP 和 AR 指标。通常情况下,AP^{50} 和 AP^{75} 分别表示阈值为 0.50 和 0.75 时的准确率,AP 指 OKS 等于 0.50、0.55、…、0.90、0.95 时的平均精度;AP^{M} 为中等尺度目标精度;AP^{L} 为较大尺度目标的精度。同理,AR 也具有相同的计算过程。

2.2 实验环境和参数

本文实验使用 Ubuntu 20.04 系统作为运行平台,采用深度学习框架 Pytorch 训练模型,编程语言主要为 Python 且通过 GPU 进行加速,GPU 型号为 NVIDIA GeForce RTX 3090,显存为 24G。

模型训练时对图像进行预处理,固定输入图像大小为 256×192,模型参数优化器为 Adam,训练批次大小为 16。网络训练共 210 个周期,学习率开始设置为 0.001,并使用余弦退火的学习率下降方式,减小到 0.000 01 不再衰减.

2.3 实验比较

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,$\mathrm{AP}=\frac{\sum_{\mathrm{m}}\sum_{\mathrm{p}}\gamma(\mathrm{OKS}_{\mathrm{p}}>t)}{\sum_{\mathrm{m}}\sum_{\mathrm{p}}1}$,其中,分母项是对“真实目标 - 预测结果”配对的全域计数。“1”作为计数单元,用于标识每一组独立的(真实目标 m、预测结果 p)配对关系,$\mathrm{OKS}\in[0,1]$ 为一个标量,$\mathrm{OKS}>t$ 表示预测正确,反之亦然,$\gamma(\cdot)$ 为预测准确性判断函数,t 为 OKS 阈值,通过实验统计得到在不

同 OKS 阈值下的识别精度。MS COCO2017 数据集 OKS 对比实验结果如表 1 所示。

由表 1 可知, 在 MS COCO2017 数据集实验中, 相较于其他模型, 本文提出的模型精度均较高。其中, 相较于 HRNet,在输入尺寸为 256×192 的情况下, 计算复杂度上升了 57.7%, 而在网络性能方面, AP 上升了 1.2%; 相较于 HRFormer, 在计算复杂度没有提升的情况下, AP^M、AP^L 分别提升了 0.3%、1.0%; 对比 VITPose 模型, 计算复杂度降低了 51.2%, 同时模型仍拥有较好的性能, AP 只降低了 0.2%; 对比 Stage Hourglass 模型、CPN 模型、Simple Baselines 模型、TokenPose 模型, AP 分别提升了 10.1%、7.0%、1.9%、0.9%, 在保持计算复杂度没有较大提升的情况下, 仍具有较高的精度, 其中表 1 中 Backbone(主干网络) 是模型的核心组成部分。GFLOPs 表示整个模型进行一次前向传播(Forward Pass)所需的总浮点运算次数。实验结果表明, 改进后的模型在标准人体姿态关键点预测性能方面优于现有相关研究。

表 1　MS COCO2017 数据集上 OKS 对比实验结果

Table 1　Experimental results of OKS comparison on MS COCO2017 datasets

Model	Backbone	Input	Parameter	GFLOPs	AP/%	AP^{50}/%	AP^{75}/%	AP^M/%	AP^L/%
CPN	ResNet	256×192	27×10^6	6.2	68.6	—	—	—	—
Simple Baseline	ResNet	256×192	60×10^6	8.9	73.7	91.9	81.1	70.3	80.0
TokenPose	HRNet	256×192	14×10^6	5.7	74.7	89.8	81.4	71.3	81.4
Stage Hourglass	StageHourglass	256×192	25×10^6	14.3	65.5	86.8	72.3	60.6	72.6
VITPose	VIT	256×192	86×10^6	17.1	75.8	90.5	83.0	—	—
HRFormer	HRFormer	256×192	43×10^6	12.2	75.6	93.6	83.6	73.2	80.1
HRNet	HRNet	256×192	29×10^6	7.1	74.4	90.5	81.9	70.8	81.0
Ours	HRNet	256×192	30×10^6	11.7	75.6	93.5	84.6	73.5	81.1

2.4　消融实验

为了验证所提出模型中各个组件的有效性, 本文设计了一系列消融实验来评估不同模块对最终性能的影响, 具体实验结果如表 2 所示。

表 2　消融研究

Table 2　Ablation studies

Model	Parameter	GFLOPs	AP/%
HRNet	28.5×10^6	7.1	74.4
GnBlock+HRNet	29.5×10^6	10.5	75.5
GnBlock+CBAM+HRNet	29.8×10^6	11.7	75.6

在 COCO 数据集上, 与 HRNet 相比较, 本文通过将 GnBlock 融入到 HRNet架构中, AP 提升了 1.1%。这证实了改进的模块可以有效提升模型性能。当融入 CBAM 后, 增加网络对重要的通道抽象特征和空间位置特征的关注, 从而获得包含更有效的信息的关键点特征, 使得 AP 提升了 0.1%, 表现出注意力机制的有效性。

2.5　实验局限

尽管本文提出的方法在人体姿态估计方面取得了一定的进展, 但仍存在一些局限性和挑战。首先, 模型对复杂场景中的姿态估计仍存在一定的局限性。递归结构需要在时间序列上逐步处理数据, 每一步都依赖于前一步的计算结果, 这导致了更高的时间复杂度。每个时间步的计算不仅涉及卷积操作, 还需要处理门控机制中的额外参数, 如更新门和重置门。这些门控单元引入了额外的参数量, 使得模型的参数总量显著增加, 从而增加了存储和计算开销。

其次, 模型的计算复杂度和资源消耗较高。虽然本文在 HRNet 的基础上进行了改进, 并通过引入 GnConv 和 CBAM 模块提升了模型性能, 但这些改进也带来了额外的计算和存储开销。相较于其他传统的姿态估计模型, 本文模型的计算复杂度与内存需求较高, 在实际应用中, 特别是资源受限的嵌入式设备或移动设备上, 如何在保持高精度的同时降低模型的计算和存储需求, 仍是一个亟待解决的问题。

此外, 训练数据的多样性和覆盖范围对模型性能的影响较大。本文使用的 COCO2017 虽然包含了大量标注数据, 但这些数据集在姿态、多样性和复杂度方面仍有一定的局限性。例如, 数据集中某些特定姿态或场景的样本量较少, 导致模型在这些情况下的泛化能力较弱。因此, 如何构建更加全面和多

样化的数据集,以提升模型在不同场景和姿态下的表现,也是未来需要研究的重要方向。

3 结束语

本文以自底向上的高分辨率姿态估计网络HRNet为基础框架,提出了一种融合递归门控卷积GnConv和卷积块注意力模块CBAM的新型姿态估计网络。通过引入GnConv,有效地捕捉图像数据中的高阶空间交互,提高特征表征能力,同时结合CBAM增强了对重要特征的关注度,提升了网络对人体姿态中细微特征的捕捉能力。实验结果表明,本文提出的模型在COCO2017数据集上的表现显著优于传统方法,证明了高阶空间交互和注意力机制在提升姿态估计精度和鲁棒性方面的有效性。未来工作中,可以进一步优化模型结构,探索更多融合不同特征的方式,以进一步提升姿态估计的性能。在处理复杂场景如遮挡问题、多目标场景、不同尺度或者视角时,可以采用例如结合轻量级卷积神经网络与图卷积网络,能够兼顾精度与计算资源需求,实现更高效的姿态估计。结合空间和通道注意力机制、深度与浅层特征的融合,提升模型在各种复杂场景下的表现,此外,引入自适应机制,使模型能够根据输入的复杂性动态调整处理策略,利用多尺度处理技术和领域适应技术,提升模型对不同尺寸和环境条件的适应能力,从而扩展其在实际应用中的适用范围。通过这些改进,未来的研究将有望显著提升姿态估计的准确性和鲁棒性,使其能够在更多复杂环境中表现出色。

参考文献:

[1] SAPP B, TASKER B. Multimodal decomposable models for human pose estimation[C]//IEEE Conference on Computer Vision and Pattern Recognition. Portland, OR, USA: IEEE, 2013: 3674-3681.

[2] DUAN H , ZHAO Y , CHEN K , *et al*. Revisiting skeleton-based action recognition[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, LA, USA: IEEE, 2022: 2959-2968.

[3] WEI W L , LIN J C , LIU T L , *et al*. Capturing humans in motion: Temporal-attentive 3D human pose and shape estimation from monocular video[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, LA, USA: IEEE, 2022: 13201-13210.

[4] DANTONE M, GALL J, LEISTNER C, *et al*. Human pose estimation using body parts dependent joint regressors[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Portland, OR, USA: IEEE, 2013: 3041-3048.

[5] LECUN Y, BOTTOU L, BENGIO Y, *et al*. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[6] 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.

[7] ZHANG K , HE P , YAO P , *et al*. DNANet: De-normalized attention based multi-resolution network for human pose estimation[EB/OL]. (2020-12-13) [2022-07-23]. https://arxiv.org/abs/1909.05090v4.

[8] ALEJANDRO N, KAIYU Y, JIA D. Stacked hourglass networks for human pose estimation[C]//European Conference on Computer Vision (ECCV). Cham: Springer, 2016: 483-499.

[9] NOH H, HONG, S, HAN B. Learning deconvolution network for semantic segmentation[C]//IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2015: 1520-1528.

[10] IGE A O, TOMAR N K, ARANUWA F O, *et al*. ConvSegNet: Automated polyp segmentation from colonoscopy using context feature refinement with multiple convolutional kernel sizes[J] IEEE Access, 2023, 11: 144082-144105.

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

[12] XU J, LIU W, XING W, *et al*. MSPENet: Multi-scale adaptive fusion and position enhancement network for human pose estimation[J]. The Visual Computer, 2023, 39(5): 2005-2019.

[13] 牛悦, 王安南, 吴胜昔. 基于注意力机制和级联金字塔网络的姿态估计 [J]. 华东理工大学学报 (自然科学版), 2023, 49(5): 724-734.

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

[15] BA J , MNIH V , KAVUKCUOGLU K . Multiple object recognition with visual attention[EB/OL]//. (2015-04-23) [2024-07-23]. https://arxiv.org /abs/1412.7755v2.

[16] WOO S H, PARK J, LEE J Y, *et al*. CBAM: Convolutional block attention module[C]//Lecture Notes in Computer Science. Munich, Germany: IEEE, 2018: 3-19.

[17] HU J, SHEN L, ALBANIE S, *et al*. Squeeze-and- excita-

tion networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 42(8): 2011-2023.

[18] WANG Z Q, XU J, LIU L, *et al*. RANet: Ranking attention network for fast video object segmentation[C]//IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, ROK: IEEE, 2019: 3977-3986.

[19] CHOLLET F. Xception: Deep learning with depthwise separable convolutions[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE, 2017: 1800-1807.

[20] RAO Y M, ZHAO W L, TANG Y S, *et al*. HorNet: Efficient high-order spatial interactions with recursive gated convolutions[EB/OL]. (2022-10-11) [2024-07-23]. https://arxiv.org/abs/2207.14284v3.

[21] VASWANI A , SHAZEER N , PARMAR N , *et al*. Attention is all you need[C]//31st Conference on Neural Information Processing Systems. Long Beach, CA, USA: NIPS, 2017: 5998-6008.

[22] YU Q , XIA Y , BAI Y , *et al*. Glance-and-gaze vision transformer[C]//35th Conference on Neural Information Processing Systems. Sydney, Australia: NeurIPS, 2021: 12992-13003.

[23] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, *et al*. Animage is worth 16×16 words: Transformers for image recognition at scale [EB/OL]. (2021-06-03) [2024-07-23]. https://arxiv.org/abs/2010.11929v2.

[24] RADFORD A, KIM J W, HALLACY C, *et al*. Learning transferable visual models from natural language supervision[EB/OL]. (2021-02-26) [2023-07-23]. https://arxiv.org/abs/2103.00020v1.

[25] HENDRYCKS D, GIMPEL K. Gaussian error linear units (GELUs) [EB/OL]. (2023-06-06) [2024-07-23]. https://arxiv.org/abs/1606.08415v5.

Pose Estimation Network Based on High-Order Spatial Interactions

HUANG Xiaoyu, CHEN Jiayi, WU Yiwei, 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: Human pose estimation is a crucial research area in computer vision. With the advancement of deep learning technologies, existing pose estimation models have achieved remarkable success in predicting human keypoints. However, when dealing with complex scenes such as severe occlusion, complex backgrounds, extreme poses, multi-scale variations, and lighting changes, these models still face challenges and their accuracy is often affected. To address this issue, this paper proposes an improved human pose estimation method based on HRNet, which significantly improves the performance of the model in complex scenes by introducing high-order spatial interaction and attention mechanisms. It employs recursive gated convolution and convolutional attention modules to enhance the model's ability to extract high-order spatial features. The experimental results show that the proposed method outperforms existing mainstream approaches on the COCO2017 dataset and achieves higher pose estimation accuracy.

Key words: pose estimation; high-resolution network; high-order spatial interaction; CBAM attention mechanism; feature extraction

（责任编辑：李娟）