doi:10.3969/j.issn.1672-1292.2024.04.005

基于注意力机制与高分辨率网络的人体姿态估计

张　铭[1],李成龙[1],高新燕[2],王鹏飞[3],张金萧[1]

(1.山东建筑大学计算机科学与技术学院,山东 济南 250000)
(2.山东华云三维科技有限公司,山东 济南 250000)
(3.中建八局第二建设有限公司,山东 济南 250000)

[**摘要**] 人体姿态估计旨在从图像或视频中精确识别关键点位置和姿态,对行为识别、人机交互等至关重要. 高分辨率网络能够从图像中提取包含多尺度信息的人体关键点特征,但主要聚焦于图像局部范围内的特征信息,难以捕捉关节间的长距离依赖,因此易受复杂背景、遮挡等因素影响,限制了准确率. 针对高分辨率网络在人体姿态估计中所面临的问题,提出了一种融合注意力机制和高分辨率网络的深度学习模块 C2F-CBAM,该模块结合了 C2F 模块和 CBAM 模块的优势,结合先进的特征提取技术和强化的注意力机制,C2F-CBAM 模块显著提高了模型在识别关键点的准确性. 此外,将 C2F-CBAM 模块嵌入到 HRNet 网络的关键位置,使得该方法能够更好地整合和综合不同尺度的特征信息. 这种融合策略不仅增强了模型对各种人体姿态和图像分辨率的适应性,还有效地处理了复杂背景和遮挡等问题. 实验结果显示,该模型在 COCO2017 验证集上相较于其他方法具有显著优势,平均精度比传统 HRNet 网络提升了 0.9,充分验证了模型的有效性和优越性.

[**关键词**] 人体姿态估计,注意力机制,高分辨率网络,C2F-CBAM 模块,关键点检测
[**中图分类号**]O643;X703　[**文献标志码**]A　[**文章编号**]1672-1292(2024)04-0046-11

Human Pose Estimation Based on Attention Mechanism and High-resolution Network

Zhang Ming[1], Li Chenglong[1], Gao Xinyan[2], Wang Pengfei[3], Zhang Jinxiao[1]

(1.School of Computer Science and Technology, Shandong Jianzhu University, Jinan 250000, China)
(2.Shandong Huayun 3D Technology Co., Ltd., Jinan 250000, China)
(The Second Construction Limited Company of China Construction Eighth Engineering Division, Jinan 250000, China)

Abstract: Human pose estimation aims to accurately identify key point positions and postures from images or videos, which is essential for behavior recognition, human-computer interaction, etc. The high-resolution network can extract the key point features of the human body containing multi-scale information from the image, but it mainly focuses on the feature information within the local range of the image, and it is difficult to capture the long-distance dependence between joints, so it is susceptible to complex background, occlusion and other factors, which limit the accuracy. In order to solve the problems faced by high-resolution networks in human pose estimation, this paper proposes a deep learning module that integrates attention mechanism and high-resolution network called C2F-CBAM, which combines the advantages of C2F module and CBAM module, and significantly improves the accuracy of the model in identifying key points by combining advanced feature extraction technology and enhanced attention mechanism. In addition, the C2F-CBAM module is embedded in the key position of the HRNet network, so that the method can better integrate and synthesize feature information at different scales, which not only enhances the adaptability of the model to various human postures and image resolutions, but also effectively deals with complex backgrounds and occlusions. Experimental results show that the proposed model has significant advantages over other methods in the COCO2017 validation set, and the average accuracy is improved by 0.9 compared with the traditional HRNet network, which fully verifies the effectiveness and superiority of the model.

Key words: human posture estimation, attention mechanisms, high-resolution networks, C2F-CBAM module, critical point detection

收稿日期:2024-05-12.
基金项目:国家自然科学基金项目(62102235)、山东省自然科学基金项目(ZR2020QF029).
通讯作者:李成龙,博士,副教授,研究方向:计算机视觉、增强现实、计算机图形学等. E-mail:lichenglong18@ sdjzu.edu.cn

二维人体姿态估计是一项富有挑战性的任务,致力于从二维图像中精确识别并定位每个人体关键点,然后将这些关键点连接成骨架,以揭示人体的姿态信息[1-4]. 然而,在实际应用中,复杂背景及遮挡问题成为制约姿态估计精度的主要障碍,也成为当前研究的热点和难点. 近年来,基于深度学习的方法已经成为该领域的研究主流,并取得了令人瞩目的进展. 这些方法大致可以归纳为两类:自下而上的方法和自上而下的方法.

自下而上方法是一种有效的人体姿态估计策略[5-7]. 如 Openpose[2]、HigherHRNet[8]等. 该类方法的优点在于可以处理图像中的所有人体,无需先检测每个人体的位置. 但自下而上方法在组合关键点信息时,可能会出现错误连接的情况,导致姿态估计结果的不准确.

自上而下方法检测人物边界框内单个人的关键点[9-11]. 人物边界框通常由对象检测器生成[12-13]. 自上而下的算法在处理图像时,首先会使用人体检测算法来检测图像中所有的人,然后针对每个人,再使用单人姿态估计方法来估计其姿态,如沙漏网络[14-15]、高分辨率网络(high-resolution network,HRNet)[16]等. 自上而下的人体姿态估计方法主要利用关键点回归和热力图回归两种方式进行关键点定位. 关键点回归[17-18]直接预测坐标位置,高效但忽略了空间信息,影响模型泛化能力. 而热力图回归[16-20]通过生成对应关键点的热力图来间接定位,提供更丰富的空间信息,对噪声、遮挡和尺度变化更鲁棒,通常能获得更高精度. HRNet 作为自上而下方法的代表,能维持高分辨率特征表示,并通过多尺度融合捕获关键信息. 但其在处理远距离关节依赖上有所不足,它更侧重于局部特征提取,而对于关键点间的长距离空间关系捕捉不够充分. 为此,本文提出了 C2F-CBAM 模块,以弥补这一缺陷并提升姿态估计性能.

C2F-CBAM 模块结合了深度卷积到两阶段金字塔网络(CSPDarknet53 to 2-Stage FPN,C2F)[20]和卷积块注意力机制(convolutional block attention module,CBAM)[21-22],旨在提升姿态估计中数据的空间特性和模型的注意力机制. C2F 模块凭借其空域特征提取能力,为姿态估计提供了坚实的基础,准确识别人体各部分. CBAM 模块通过通道注意力机制[23]和空间注意力机制[24],增强了模型对姿态关键点的敏感度,理解全局上下文和局部结构. 本文将 C2F-CBAM 嵌入到 HRNet 中,保留了 HRNet 的高分辨率和多尺度融合优势,并显著增强了模型对长距离关节依赖的建模能力,使模型在复杂场景下能更准确地捕捉关键点间的空间关系,实现精确稳定的姿态估计.

1　相关工作

近年来,随着深度学习技术的持续进步,人体姿态估计领域中的自下而上方法和自上而下方法均获得了显著的改进. 为了克服复杂背景和遮挡等难题带来的挑战,研究者们致力于研究更加强大和鲁棒的人体关键点检测算法,通过优化网络结构、增强特征提取能力以及引入多尺度信息融合等技术,力求从图像或视频中精确识别和定位人体关键点.

OpenPose 作为自下而上策略的代表方法,用自下而上策略,通过两个分支网络预测人体姿态:一个分支预测关节位置,另一个预测关节间的关联(部分亲和场),确保关键点组合正确. 自上而下的姿态估计方法往往依赖于人物检测器的准确性,因为姿态估计是在检测到的人物区域上进行的. 因此,人物定位错误或边界框预测不准确会影响姿态提取的性能[25]. 而 HigherHRNet 则通过高分辨率特征金字塔学习尺度感知表示,结合多分辨率监督和聚合,有效解决了自下而上多人姿态估计中的尺度变化挑战,并能更精确地定位关键点,特别是在处理小人物时表现更佳.

自上而下的人体姿态估计是策略的经典方法之一[25-26],它采用多阶段骨干网络提取图像特征,并通过关键点回归网络检测每个边界框内的关键点. 沙漏网络则以其独特的结构捕获并整合多尺度特征信息,通过编码阶段捕获全局特征,解码阶段结合多尺度特征信息得到精细的局部特征表示,进一步提升了姿态估计的准确性和精细度. HRNet 从高分辨率子网开始,逐步引入多个并行的多分辨率子网. 这些子网处理不同尺度的特征,并通过跨尺度连接进行信息交换和融合. 这种结构允许 HRNet 同时捕获多尺度信息,从而深入理解图像内容. 通过并行子网间的信息融合,HRNet 有效结合了低分辨率的强语义信息和高分辨率的精确位置信息. YOLO-Pose[27]在保持 YOLOv5[28]快速和准确的目标检测能力的同时,增加了一个并行的关键点检测分支,这个分支负责在每个人体边界框内精确地定位关键点,通过这种方式,YOLO-

Pose 能够在一个统一的框架内同时解决人体检测和姿态估计两个问题,从而提高了整体的处理效率和准确性. 高分辨率(high resolution transformer,HRFormer)[29]利用多分辨率架构与局部窗口自注意力机制,可以有效实现高分辨率特征表示,同时保持低内存和低计算成本. 然而,该方法中的上采样操作会导致空间语义信息的丢失,这是一个挑战.

在计算机视觉任务中,注意力机制已成为提升模型性能的关键. SENet[30]和 ECA[31]作为通道注意力的代表,通过聚合全局上下文信息并对特征通道进行加权,有效凸显了关键特征并抑制了冗余信息. 同时,非局部神经网络[32]及其改进版本通过捕捉特征间的长距离依赖,增强了模型对全局信息的感知能力. CBAM[22]等方法则进一步融合了通道注意力和空间注意力,从多维度优化特征表达,显著提升了模型的性能和泛化能力. 这些注意力机制使模型在处理视觉任务时更加聚焦,从而取得了更好的效果.

2　基本方法

本文提出了一种网络架构——HRNet-C2F-CBAM,如图 1 所示. 该网络融合了注意力机制与 HRNet 网络的优势. HRNet-C2F-CBAM 的网络结构通过整合不同尺度的特征信息能够有效地应对背景噪声、遮挡以及自遮挡等复杂场景,显著提升了模型的性能.

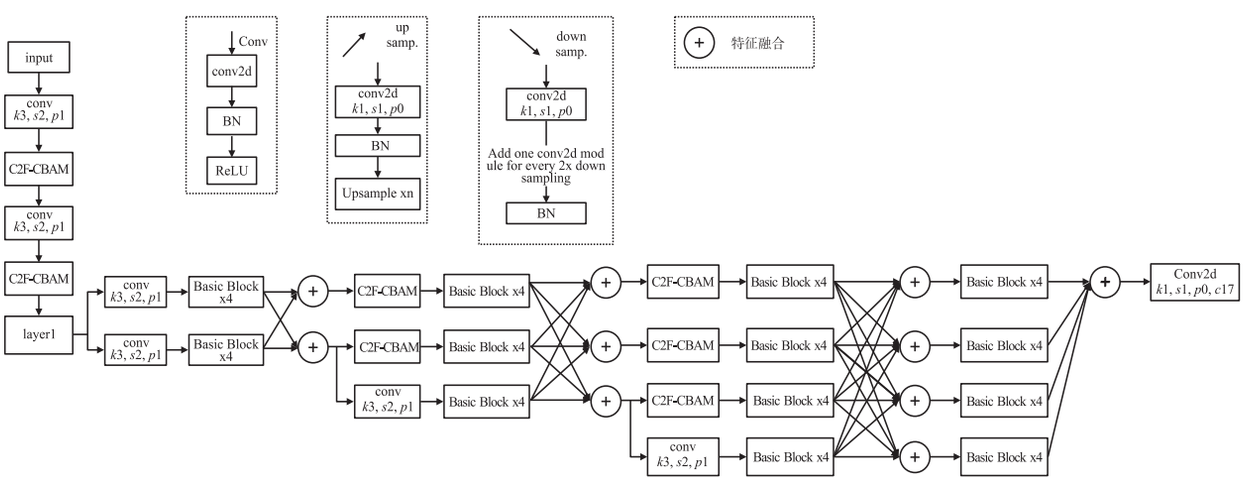


图 1　HRNet-C2F-CBAM 网络图

Fig. 1　Diagram of HRNet-C2F-CBAM network

2.1　HRNet 网络

HRNet 网络作为一种深度学习网络结构,其核心优势在于能够在整个处理过程中维护高分辨率的表示,这得益于其精心设计的多个 Stage 模块和过渡模块. 其中每个 Stage 模块作为子网络专注于处理不同分辨率的特征,这些并行连接的 Stage 模块确保网络能同时利用多个尺度的信息,在 Stage 模块内,通过堆叠构建块,网络逐步提取、转换并增强特征表示. 过渡模块连接 HRNet 中相邻的 Stage 模块,实现跨分辨率信息交互,它通过采样调整特征图分辨率,融合不同分辨率的特征,结合语义和位置信息,生成更准确的特征表示. 然而,HRNet 网络难有效地捕捉人体关键点之间的长距离空间依赖关系,这导致在复杂背景、人体自遮挡或人体间遮挡等情况下姿态估计的准确率受到限制.

2.2　C2F 与 CBAM 的结合和新模块的设计原理

为了解决复杂背景、人体自遮挡或人体间遮挡等问题,本文提出了 C2F-CBAM 模块,并将其融合到 HRNet 网络中. C2F-CBAM 模块的设计旨在更好地整合和综合不同尺度的特征信息,从而提高姿态估计精度. 通过引入这个模块,期望 HRNet 网络能够在处理具有挑战性的姿态估计任务时表现出更高的准确性和鲁棒性.

2.2.1　C2F 模块

C2F 模块的主要功能是对输入数据进行空间特征提取. 通过一系列的卷积操作,它可以在不同的空间位置上捕捉和识别图像的关键特征. C2F 模块通常包含多个卷积层和非线性激活函数,如图 2 所示. 这

些卷积层用于提取空间特征,并通过引入非线性激活函数,增强模型的表达能力. 此外,C2F 模块还可以包含批归一化层和池化层,用于加速训练和增强特征的鲁棒性.

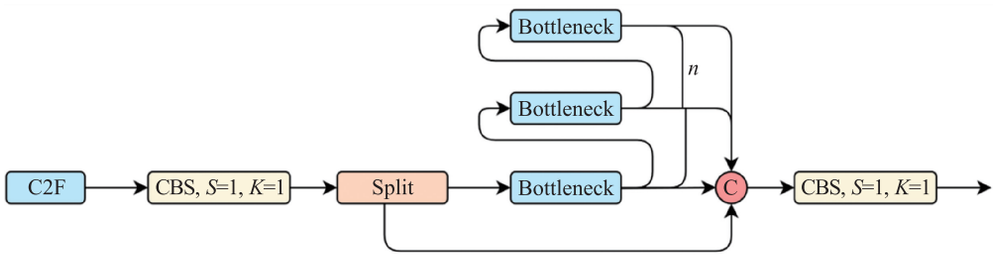


图 2　C2F 网络结构图

Fig. 2　Diagram of C2F network structure

2.2.2　CBAM 模块

CBAM 模块是一种包含通道注意力和空间注意力机制的注意力模块,如图 3 所示.

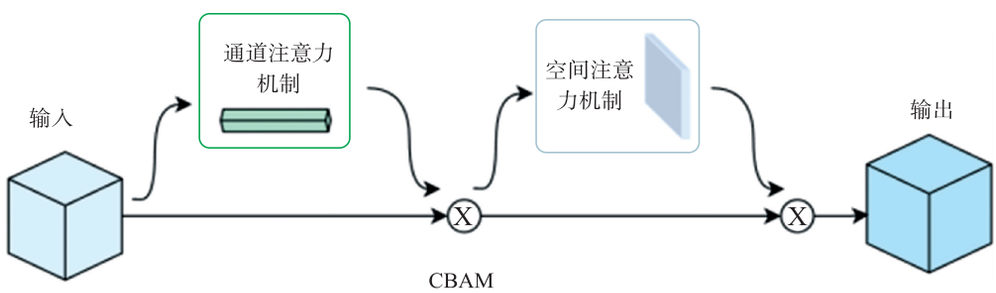


图 3　CBAM 网络结构图

Fig. 3　Diagram of CBAM network structure

通道注意力为通过通道注意力,模型可以识别出重要特征,并赋予它们更大的权重. 通道注意力的计算可以表示为

$$Ac=\sigma(Wc\times\mathrm{AvgPool}(X)+Wc\times\mathrm{MaxPool}(X)). \tag{1}$$

其中,Ac 表示通道注意力权重,σ 是激活函数(通常是 Sigmoid),Wc 是学习的权重参数,×表示卷积操作,AvgPool 和 MaxPool 分别表示平均池化和最大池化操作.

空间注意力机制关注于特征图的空间位置信息. 可以捕捉到图像中不同区域的重要性,使模型能够定位并关注到关键的部位或区域. 空间注意力的计算可以表示为

$$As=\sigma(Ws\times\mathrm{AvgPool}(X)+Ws\times\mathrm{MaxPool}(X)). \tag{2}$$

其中,As 表示空间注意力权重,σ 是激活函数,Ws 是学习的权重参数,×表示卷积操作.

2.2.3　C2F-CBAM 模块

由于 C2F 模块中的 Bottleneck 不能主动学习特征图中的重要区域,且计算复杂度较高,导致模型训练和推断的速度较慢. 本文提出用 CBAM 模块替换 C2F 模块中的 Bottleneck,将 C2F 与 CBAM 结合后形成的新模块称为 C2F-CBAM,如图 4 所示. C2F-CBAM 是一种结合了通道注意力、空间注意力和空域信息的深度学习模块. 该模块融合 C2F 的多尺度特征提取与 CBAM 的注意力机制,旨在通过增强模型的特征提取能力及对全局和局部信息的理解,提升姿态估计任务的性能.

原始的输入特征图送入 C2F 模块,假设输入特征图 X,$X\in h\times w\times c$ 其中 h 和 w 分别表示特征图的高度和宽度,c 表示通道数. C2F 模块包含一系列的卷积操作,用于提取空域特征. 卷积操作可表示为

$$F_{c2}=\sum_i\sum_j K_{i,j}\times X_{i,j}. \tag{3}$$

其中,F_{c2}为卷积后的结果,K 为卷积操作,X 为输入特征图.

使用 CBAM 注意力机制替代 Bottleneck,将通道注意力图和空间注意力图与中间特征图进行逐元素相乘,对特征图进行加权处理. 这个过程中,重要的特征通道和空间位置被赋予更高的权重,以增强模型对关键特征的关注.

$$z = F_{c2} \times A_c \times A_s. \tag{4}$$

输出特征图经过 CBAM 注意力机制加权处理后的特征图作为 C2F 模块的输出,被送入 HRNet 网络中继续进行处理

$$\text{output} = \text{C2F-CBAM}(X) = Z. \tag{5}$$

其中,C2F-CBAM 表示结合了 C2F 和 CBAM 的模块.

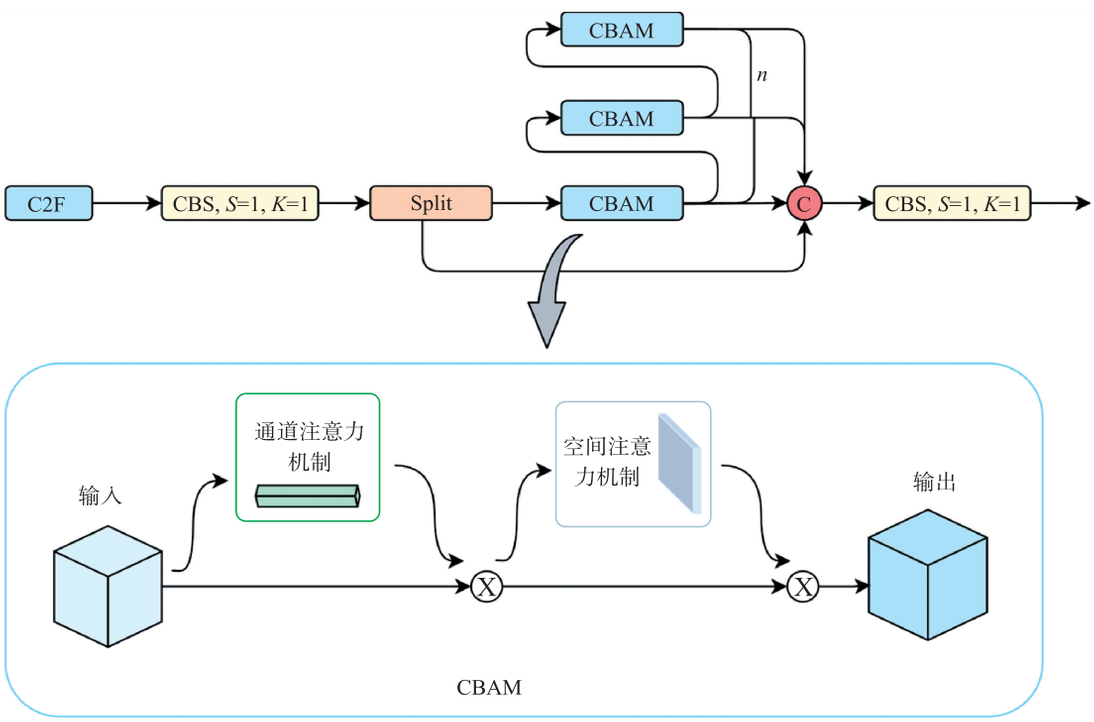


图 4　C2F-CBAM 模块结构图

Fig. 4　Diagram of C2F-CBAM module structure

2.3　模块融合

2.3.1　C2F-CBAM 模块融合到卷积层

将 C2F-CBAM 模块融合到卷积层,如图 5 所示. C2F-CBAM 模块结合了 C2F 模块的空间特征提取能力和 CBAM 模块的注意力机制,能够更有效地捕捉和识别图像中的关键特征,并提升模型对重要特征区域的关注度.

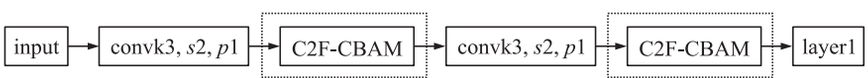


图 5　C2F-CBAM 模块融合到卷积层

Fig. 5　The C2F-CBAM module is fused to the convolutional layer

2.3.2　C2F-CBAM 模块融合到多尺度分支

在多尺度分支引入 C2F-CBAM 模块:在图 1 的基础上将 C2F-CBAM 模块引入到 HRNet 的多尺度分支后,这样可以确保 C2F-CBAM 模块能够在多尺度特征处理中发挥作用.

将 C2F-CBAM 模块的输出与多尺度分支的特征图进行连接. 通过拼接操作实现,将 C2F-CBAM 模块的输出特征图与多尺度分支中相应尺度的特征图在通道维度上进行拼接.拼接完成后,进行特征融合和调整操作,确保后续网络层处理新的特征表示. 包括使用卷积层进行调整,减少特征图的通道数,以及进行批归一化和激活函数等操作,增加模型的非线性表达能力.经过特征融合与调整后,继续将特征图送入 HRNet 的多尺度分支中进行进一步的处理. C2F-CBAM 模块提取的特征与其他尺度的特征一起进行多尺度的特征融合和交互,以提高模型对全局和局部信息的理解.

3　实验分析

3.1　实验环境

本文实验平台系统为 Windows 10 专业版,处理器为 Intel(R)Xeon(R)Gold 5218R CPU@2.10 GHz,GPU 为 NVIDIA RTX A6000.深度学习框架为 2.0.1 版本的 Pytorch,编程语言为 Python3.9,实验中采用了多种数据增强技术. 在优化模型参数方面,选用了 Adamw 优化器[33],将基本学习率设为 1E-3,并在第 170 和 200 个训练周期时,分别将其降低至 1E-4 和 1E-5,整个训练过程在 210 个训练周期内完成.

3.2　数据集介绍

本研究使用数据集 COCO[34],其包含了 200 000 张图像和 250 000 个使用 17 个关键点标记的 Person 实例. 选用 COCO Train 2017 数据集进行模型的训练,该子集包含了 57K 图像和 15 万个 Person 实例. 在训练完成后,在 val2017 数据集上对模型进行了评估,val2017 数据集包含了 5 000 张图像,提供了足够的数据来学习并优化模型.

3.3　评价指标介绍

为了评估模型的性能,本文使用了基于对象关键点相似性(*OKS*)的标准评估度量.

$$OKS=\frac{\sum_i[e^{-d_i^2/2s^2k_i^2}\cdot\delta(v_i>0)]}{\sum_i[\delta(v_i>0)]}. \tag{6}$$

其中,d_i 是模型预测的关键点与 GT 之间的欧氏距离,v_i 代表第 i 个关键点的可见性,由 GT 提供,s 为目标面积的平方根,目标面积为分割面积,该数据在 COCO 数据集标注信息中均有提供. *OKS* 的值通常在 0 到 1 之间,表示匹配的质量,其中 1 表示完美匹配.

该方法是一种常用的姿态估计任务评估方法,主要衡量预测的关键点与真实关键点之间的相似性. 报告了标准平均准确率和召回率得分,包括 *AP*50(即 *OKS*=0.50 时的平均精度)、*AP*75,以及 10 个位置的 *AP* 得分的平均值(从 0.50 到 0.95). 另外,还使用了 *APM*(中型对象平均精度)和 *APL*(大型对象平均精度)两个指标,以更全面地评估模型在不同大小的目标上的性能. 最后,计算了 *AR*(平均召回率)在 *OKS*=0.50 到 0.95 的范围内的得分. 通过评估指标,可更全面地了解本文的模型在姿态估计任务上的性能,并与其他方法进行比较.

3.4　定量分析

在训练过程中,先对人体检测框进行了调整,再从图像中裁剪出来,将其高度或宽度扩展至固定的长宽比. 在测试过程中采用两阶段自顶向下的范式进行姿态估计,利用人物检测器来识别图像中的人物实例. 其中,选择 YOLOv8[35]作为人物检测器,在人物实例被检测出来后,再进行关键点的预测,以完成整个姿态估计的过程.

将本文方法与其他先进方法进行比较,结果如表 1 所示.

表 1　HRNet-C2F-CBAM 算法与其他算法在数据集 MS COCO 的实验结果对比

Table 1　Comparison of the experimental results of the HRNet-C2F-CBAM algorithm with other algorithms on the dataset MS COCO

方法	主干网络	是否预训练	输入大小	*AP*	*AP*50	*AP*75	*APM*	*APL*	*AR*
8-stage Hourglass	8-stage Hourglass	否	256 x 192	66.9	—	—	—	—	—
CPN	ResNet-50	是	256 x 192	68.9	—	—	—	—	—
CPN+OHKM	ResNet-50	是	256 x 192	69.4	—	—	—	—	—
SimpleBaseline	ResNet-50	是	256 x 192	70.4	88.6	78.3	67.1	77.2	77.2
SimpleBaseline	ResNet-101	是	256 x 192	71.4	89.3	79.3	68.1	78.1	77.1
SimpleBaseline	ResNet-152	是	256 x 192	72.0	89.3	79.8	68.7	78.9	77.8
PRTR	HRNet-W32	是	512 x 384	73.3	89.2	79.9	69.0	80.9	80.2
HRFormer-T	HRNet-W32	是	256 x 192	70.9	89.0	78.4	67.2	77.8	76.6
HRNet	HRNet-W32	否	256 x 192	73.4	89.5	80.7	70.2	80.1	78.9
HRNet-C2F-CBAM	HRNet-W32	否	256 x 192	74.0	92.4	81.5	71.1	78.4	77.0
HRNet	HRNet-W32	是	256 x 192	74.4	90.5	81.9	70.8	81.0	79.8
HRNet-C2F	HRNet-W32	是	256 x 192	74.5	92.5	82.1	71.2	79.3	77.4
HRNet-C2F-CBAM	HRNet-W32	是	256 x 192	75.3	93.4	82.2	72.3	80.0	78.4

从实验结果中,HRNet 基线方法仅使用单尺度测试,也能超越使用多尺度测试的 Hourglass 方法. 而本文进一步提出的 HRNet-C2F-CBAM 算法,在性能上有显著提升,HRNet-C2F-CBAM 算法在没有使用预训练模型的情况下,已达到 74.0,比 HRNet 高出 0.6. 若采用预训练模型,HRNet-C2F-CBAM 算法的性能还能进一步提升至 75.3,比 HRNet 高出 0.9. 这不仅超越了 HRNet 本身,还比 HRFormer(W32 网络,输入为 256X192)高出 4.4,比 PRTR[35] 网络高出 2.0,明显优于大部分现有的自上向下方法. 实验结果证明了 HRNet-C2F-CBAM 算法在关键点检测精度和误差方面均表现更佳. 表明 HRNet-C2F-CBAM 算法通过增强模型的特征表示和引入注意力机制,有效地提升了姿态估计的性能.

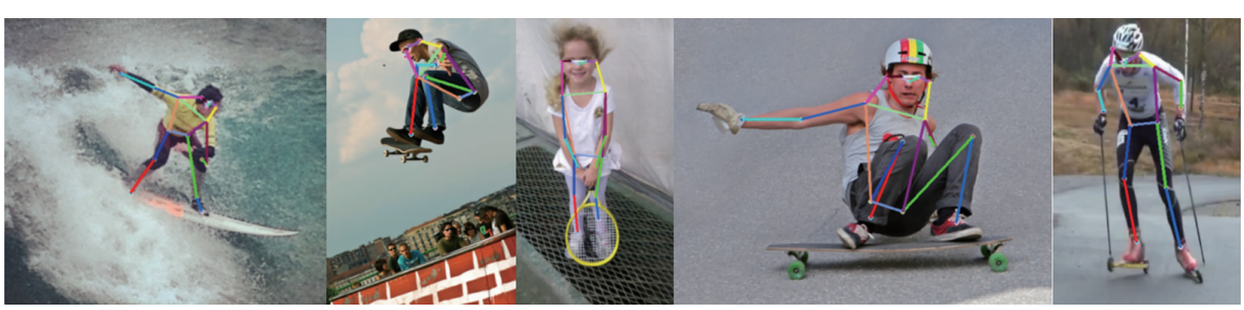
3.5　定性分析

HRNet-C2F-CBAM 算法的可视化结果如图 6 所示,证明了该算法在提升模型对全局和局部信息理解能力方面的有效性. 通过引入 HRNet 的多分辨率并行结构和 C2F-CBAM 模块,模型在处理复杂背景、人体自遮挡以及人体间遮挡等具有挑战性的场景时,展现出了更优越的性能.

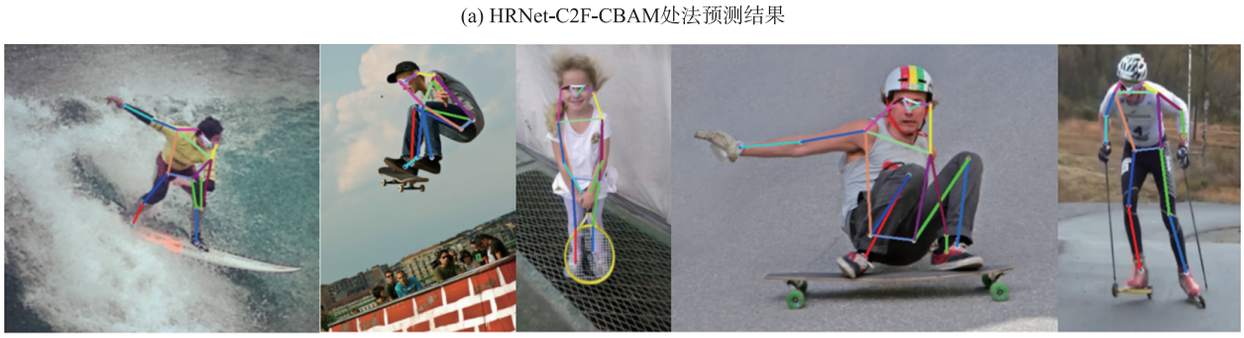
图 6　HRNet-C2F-CBAM 模型的预测结果

Fig. 6　Predictive results of the HRNet-C2F-CBAM model

算法 HRNet-C2F-CBAM 与 HRNet 模型在可视化结果上的直接比较如图 7 所示. 在该组对比图中,可清晰地看到两者在关键点检测准确性上的差异. 在图 7(a)中,HRNet-C2F-CBAM 算法在关键点检测方面



(a) HRNet-C2F-CBAM处法预测结果



(b) HRNet算法预测结果

图 7　两种算法预测结果

Fig. 7　Predictive results of two algorithms

表现出了更高的精确度和细致性,预测的关键点位置与真实位置非常接近,偏差很小. 在图 7(b)中,HRNet 模型也能够进行一定程度的关键点检测,但从可视化结果来看,其预测的关键点位置相对不够精确,存在一定的偏差. 由于 HRNet 模型主要聚焦于图像局部范围内的特征信息,难以捕捉关节间的长距离依赖的局限性所致. 通过直接对比图 7(a)和图 7(b)的预测结果,可直观地感受到 HRNet-C2F-CBAM 模型相较于基础 HRNet 模型在关键点检测精度上的显著提升.

3.6　消融实验

在 HRNet 网络上分别添加 C2F 模块、CBAM 模块以及组合使用 C2F-CBAM 模块进行对比实验,探讨模块对姿态估计性能的影响. 表 2 详细分析了 HRNet 基线模型以及在其基础上单独或联合引入这些模块的对比结果. 实验数据显示,相较于原始 HRNet,单独添加 C2F 模块或 CBAM 模块均带来了一定程度上的性能提升. 但将 C2F 模块与 CBAM 模块结合使用将得到最显著的提升. 这些实验结果清晰地表明,C2F-CBAM 模块的加入为 HRNet 模型带来了可观的性能提升. 由于 C2F 模块增强了特征的跨通道交互,而 CBAM 模块则通过注意力机制优化了特征的空间和通道权重分配. 当这两个模块联合使用时,它们相互补充,进一步提升模型的特征表示能力和关键点检测精度.

表 2　消融实验的对比结果

Table 2　Comparative results of ablation experiments

方法	主干网络	是否预训练	输入大小	*AP*	*AP*50	*AP*75	*APM*	*APL*	*AR*
HRNet	HRNet-W32	否	256 x 192	73.4	89.5	80.7	70.2	80.1	78.9
HRNet-C2F	HRNet-W32	否	512 x 384	70.0	90.3	77.7	68.2	73.6	75.0
HRNet-CBAM	HRNet-W32	否	256 x 192	72.0	92.4	80.3	70.2	77.2	76.0
HRNet-C2F-CBAM	HRNet-W32	否	256 x 192	74.0	92.4	81.5	71.1	78.4	77.0
HRNet	HRNet-W32	是	256 x 192	74.4	90.5	81.9	70.8	81.0	79.8
HRNet-C2F	HRNet-W32	是	256 x 192	74.5	92.5	80.2	72.0	79.0	77.4
HRNet-CBAM	HRNet-W32	是	256 x 192	74.2	92.4	81.4	71.2	78.7	77.2
HRNet-C2F-CBAM	HRNet-W32	是	256 x 192	75.3	93.4	82.2	72.3	80.0	78.4

图 8、图 9 展示了两组对比图,直观的可视化图片展示了各模块在添加到 HRNet 网络后产生的效果,用于比较 HRNet-C2F-CBAM 的优势. 在这两组对比图中,我们可以清晰地看到,无论是否使用预训练模

(a) HRNet-C2F-CBAM算法的预测结果

(b) HRNet-C2F算法的预测结果

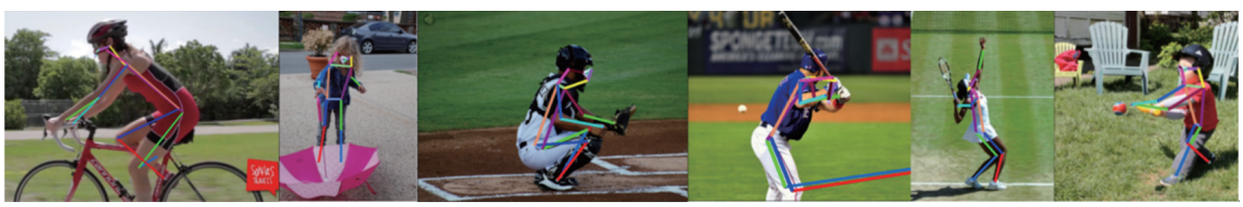
(c) HRNet-CBAM算法的预测结果

图 8　使用 ImageNet 文件夹中预训练模型产生的 3 种算法预测结果

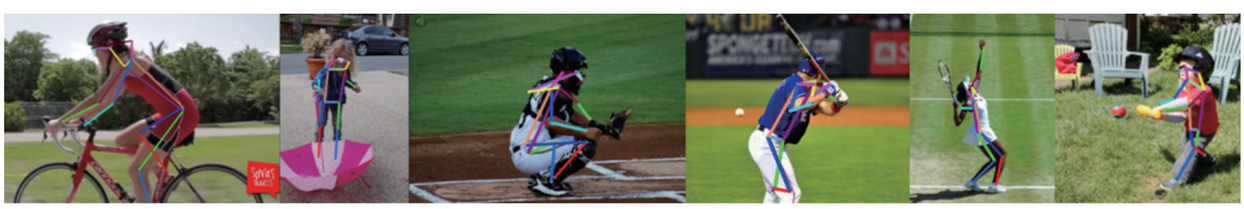
Fig. 8　Predictive results of three algorithms by using the pre-trained model in the ImageNet folder

型,HRNet-C2F-CBAM 算法都表现出了卓越的性能. 进一步验证了 HRNet-C2F-CBAM 算法在关键点检测任务中的优势和有效性.

(a) HRNet-C2F-CBAM算法的预测结果



(b) HRNet-C2F算法的预测结果



(c) HRNet-CBAM算法的预测结果

图 9　未使用预训练模型产生的 3 种算法的预测结果

Fig. 9　Predictive results of three algorithms without using a pre-trained model

4　结论

本文提出了 HRNet-C2F-CBAM 算法,该算法巧妙地融合了 C2F 模块与 CBAM 模块. C2F 模块通过其独特的设计,能够提取丰富的空域信息,能更好地捕获图像的结构和特征. 而 CBAM 模块的引入,则进一步增强了模型对特征的选择和权重分配能力,将空域信息和注意力机制有机结合,算法为模型提供了更为丰富的信息表示能力,捕捉关节间的长距离依赖信息,使其在复杂的图像场景中也能表现出色,能够有效处理背景噪声、遮挡和自遮挡等问题,充分证明了 HRNet-C2F-CBAM 算法的有效性和实用性,为姿态估计领域的研究提供了新的思路和方法.

[参考文献](References)

[1] LI K,WANG S J,ZHANG X,et al. Pose recognition with cascade transformers[C]//IEEE Conference on Computer Vision and Pattern Recognition. Nashville,TN,USA,2021.

[2] CAO Z,SIMO T,WEI S,et al. Realtime multi-person 2d pose estimation using part affinity fields[C]//IEEE Conference On Computer Vision And Pattern Recognition. Honolulu,HI,USA,2017.

[3] KOCABAS M,KARAGOZ S,AKBAS E. Multiposenet:Fastmulti-person pose estimation using pose residual network[C]// European Conference on Computer Vision. Munich,Germany,2018.

[4] PAPANDREOU G,ZHU T,GIDARIS S,et al. Personlab:Person pose estimation and instance segmentation with a bottom-up, part-based,geometric embedding model[C]//European Conference on Computer Vision. Munich,Germany,2018.

[5] NEWELL A,HUANG Z A,DENG J. Associative embedding:End-to-end learning for joint detection and grouping[C]// NeurIPS. Long Beach,CA,USA,2017.

[6] INSAFUTDINOV E,PISHCHULIN L,ANDRES B,et al. Deepercut:A deeper,stronger,and faster multi-person pose estimation model[C]//European Conference on Computer Vision,Amsterdam,The Netherlands,Amsterdam. The Netherlands,2016.

[7] 孔英会,秦胤峰,张珂. 深度学习二维人体姿态估计方法综述[J].中国图象图形学报,2023,28(7):1965-1989.

[8] CHENG B W,XIAO B,et al. HigherHRNet:Scale-aware representation learning for bottom-up human pose estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle,WA,USA,2020.

[9] 邹宇翔,何宁,郭宇昕,等. 基于深度学习的人体姿态估计综述[C]//中国计算机用户协会网络应用分会 2023 年第二十七届网络新技术与应用年会. 镇江,江苏,2023.

[10] XIAO B,WU H P,WEI Y C. Simple baselines for human pose estimation and tracking[C]//European Conference on Computer Vision. Munich,Germany,2018.

[11] WANG J D,SUN K,CHENG T H,et al. Deep high-resolution representation learning for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligenc,2021,43(10):3349-3361.

[12] CHENG B W,WEI Y C,FERIS R,et al. Decoupled classification refinement:Hard false positive suppression for object detection[J]. arXiv Preprint arXiv:1810.04002,2018.

[13] CHENG B W,WEI Y C,SHI H H,et al. Revisiting rcnn:On awakening the classification power of faster rcnn[C]//European Conference on Computer Vision. Munich,Germany,2018.

[14] NEWELL A,YANG K,DENG J. Stacked hourglass networks for human poseestimation[C]//European Conference on Computer Vision. Amsterdam,The Netherlands,2016.

[15] CHU X,YANG W,OUYANG W,et al. Multi-context attention for human pose estimation[C]//IEEE Conference on Computer Vision and Pattern Recognition. Honolulu,HI,USA,2017.

[16] SUN K,XIAO B,LIU Det al. Deep high-resolution representation learning for human pose estimation[C]//IEEE Conference on Computer Vision and Pattern Recognition. Long Beach,CA,USA,2019.

[17] CARREIRA J,AGRAWAL P,FRAGKIADAKI K. Human pose estimation with iterative error feedback[C]//IEEE Conference On Computer Vision And Pattern Recognition. Las Vegas,NV,USA,2016.

[18] TOSHEV A,SZEGEDY C. Deeppose:Human pose estimation via deep neural networks[C]//IEEE Conference on Computer Vision and Pattern Recognition. Columbus,OH,USA,2014.

[19] CHU X,OUYANG W,LI H,et al. Structured feature learning for pose estimation[C]//IEEE Conference on Computer Vision and Pattern Recognition. LasVegas,NV,USA,2016.

[20] YANG W,OUYANG W,LI H,et al. End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation[C]//IEEE Conference on Computer Vision And Pattern Recognition,Las Vegas,NV,USA,2016.

[21] TOMPSON J,GOROSHIN R,JAIN A,et al. Efficient object localization using convolutional networks[C]//IEEE Conference on Computer Vision and Pattern Recognition,Boston,MA,USA,2015.

[22] WOO,SANGHYUN,et al. Cbam:Convolutional block attention module[C]//European Conference on Computer Vision. Munich,Germany,2018.

[23] JIE H,SHEN L,SUN G. Squeeze-and-excitation networks[C]//IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City,UT,USA,2018

[24] JADERBERG MAX,SIMONYAN K,ZISSERMAN A,et al. Spatial transformer networks[J]. Advances in Neural Information Processing Systems,2015,2:2017-2025.

[25] FANG H S,XIE S Q,TAI Y W,et al. Rmpe:Regional multi-person pose estimation[C]//European Conference on Computer Vision. Honolulu,HI,USA,2017.

[26] YUAN Y,FU R,HUANG L,et al. High-resolution transformer for dense prediction[J]. arXiv Preprint arXiv:2110.09408,2021.

[27] DEBAPRIYA M J,NAGORI S,MATHEW M,et al. Yolo-pose:Enhancing yolo for multi person pose estimation using object keypoint similarity loss[C]//IEEE Conference on Computer Vision and Pattern Recognition. New Orleans,LA,USA,2022.

[28] ZHU X,LVU S,WANG X,et al. TPH-YOLOv5:Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios[C]//IEEE Conference on Computer Vision Recognition. Nashville,TN,USA,2021.

[29] YUAN Y H,FU R,HUANG L,et al. HR Former:High-Resolution transformer for dense prediction[J]. arXiv Preprint arXiv:2110.09408,2021.

[30] HU J,SHEN L,SUN G. Squeeze-and-excitation networks[C]//IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City,UT,USA,2018.

[31] WANG Q,WU B,ZHU P,et al. ECA-Net:Efficient channel attention for deep convolutional neural networks[C]//IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City,UT,USA,2020.

[32] GIRSHICK R, GUPTA A, et al. Non-local neural networks[C]//IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA, 2018.

[33] LOSHCHILOV I, HUTTER F. Decoupled weight decay regularization[J]. arXiv Preprint arXiv:1711.05101, 2017.

[34] LIN T, MAIRE M, BELONGIE S, et al. Microsoft COCO: Com-mon objects in context[C]//European Conference on Computer Vision. Zurich, Switzerland, 2014.

[35] VARGHESE R, SAMBATH. YOLOv8: A novel object detection algorithm with enhance performance and robustness[C]// 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems. Chennai, India, 2024.

[责任编辑:陈　庆]