

文章编号: 2096-1618(2025)02-0157-06

基于改进双注意力机制的轻量型人体姿态检测算法

唐芷宣^{1,2}, 高瑜翔^{1,2}

(1. 成都信息工程大学通信工程学院, 四川 成都 610225; 2. 气象信息与信号处理四川省高校重点实验室, 四川 成都 610225)

摘要:为提高轻量化人体姿态检测算法的准确率,提出一种新的双注意力机制的轻量化高分辨率人体姿态检测网络 ES-LHRNet。借鉴 HRNet(high-resolution network)的框架,采用稠密连接网络和堆叠的轻量化倒残差结构进行特征提取,并提出一种新的融合通道注意力和空间注意力机制的双注意力机制捕获位置信息和通道信息提升算法准确率。相比于 HRNet,ES-LHRNet 参数量减少 86%,运算复杂度下降 73%,并且提出的 ESAM 使在数据集 MS COCO 2017 上的检测结果平均精度提升了1.6个百分点。

关键词:人体关键点检测;高分辨率网络;注意力机制;特征复用;轻量型网络

中图分类号:TP391

文献标志码:A

doi:10.16836/j.cnki.jcuit.2025.02.006

0 引言

人体姿态检测广泛应用于动画制作、无人驾驶以及智能交互等领域。人体姿态检测任务是通过检测目标中的人体关键点的位置信息,并对信息进行处理得到人体姿态的检测结果。自 Toshev 等^[1]开创将深度学习用于人体关键点检测的研究并将关键点检测建模为关键点的坐标的回归问题以来,大批研究者将姿态识别的研究重点从基于图结构等传统方法转移到深度学习算法上。深度学习处理人体姿态识别的主要过程仍然是利用卷积不断学习图片的特征,最后得到像素点上关键点的预测值。在基于深度学习的姿态检测中,研究者完成了从单人到多人实时检测的发展;提出自顶向下以及自底向上的姿态检测算法实现不同的需求,自顶向下有着更高的关键点检测精度,而自底向上有着更快的检测速率。同时,为深度学习算法更便利地应用于嵌入式设备,研究者们也追求网络的轻量化。

目前的姿态检测研究中,关键点检测的准确率仍是一个待完善的问题,不仅如此,还要求一个高性能姿态检测网络拥有更少的参数量和计算量。为提高检测精度,Wei 等^[2]在提出的 CPM(convolutional pose machines)方法中首次采用以热力图回归关键点的预测值的方法得到检测结果,解决实际图片中的人体关键点并不能被某一个像素点指定给算法带来的误差,热力图的方式使姿态检测网络的准确率有较大的提升。以后的研究沿用这个思想,但 CPM 作为一个端到端的多人检测网络,仍然存在检测效果不稳定的问题。为

继续提高检测精度,Cao 等^[3]提出级联金字塔网络结构和始终保持主干高分辨率网络的网络结构来替代传统的从高到低分辨率的网络结构。Chen 等^[4]提出级联金字塔网络(cascaded pyramid network,CPN)提高了关键点被遮挡时的检测性能。该网络分为两个步骤:GlobalNet 主要负责检测容易检测和需要更深层的语义信息来解决的较难检测的关键点;RefineNet 主要解决需要通过训练损失的更难或者不可见关键点的检测。Sun 等^[5]提出一种始终保持主干网络为高分辨率进行特征提取的人体姿态检测网络 HRNet。该网络结构使模型能不断地学习不同分辨率率大小的特征,准确率也在 2019 年达到最高,但这类网络也存在参数量大网络结构复杂导致训练较困难和对硬件要求较高的问题。为得到一个轻量化的姿态检测网络,Yu 等^[6]提出基于 HRNet 和 ShuffleNet V2^[7]的轻量化网络 Lite-HRNet。

从 Lite-HRNet 中可以看出减少特征提取模块以及使用 Shufflenet V2 的特征提取模块来替代 HRNet 中的残差模块减少了大量的参数量和计算复杂度,但同时也削弱了算法的检测精度。本文在特征提取上,借鉴 ShuffleNet V2 的特征提取模块对 HRNet 轻量化的同时,引入 DenseNet 中提取的特征提取模块替换网络的瓶颈结构,利用 DenseNet 的特征复用来避免仅仅使用 ShuffleNet V2 带来的损失。同时提出新的融合空间和通道的双注意力机制来提升准确率。

1 方法

本文提出的 ES-LHRNet 是通过引入轻量化的倒

残差结构和稠密连接网络以及双注意力机制对 HRNet-W32 进行改进的,整体特征图的提取过程以及网络处理的整体架构如图 1 所示。HRNet^[5] 是自上向下的 2D 人体姿态检测网络,核心思想是在保持高分辨率特征的同时下采样获得高语义信息并从不同尺寸特征中融合信息,不断丰富高分辨率特征的信息。在原 HRNet 网络中,特征提取主要是由 ResNet 中提出的 bottleneck 结构和 basicblock 结构完成,在特征提取的过程中保持 64×48 高分辨率特征图,不断下采样增加通道数获取更多特征信息的同时,需要不断将不同尺度的信息融合在高分辨率特征中以弥补原始图像信息因为特征感受野的扩大而造成的损失,获得更多的空间结构关系以及语义层面的信息,以达到改善关键点

定位的效果^[8]。本文方法沿用 HRNet 网络的整体架构和思路,将预处理后的图片输入网络中,通过两次 conv2d+rule+BN 操作将特征图大小处理为 64×48 ,通道数增加至 64;在 stage1 中采用改进的稠密连接网络结构,使得保持输出特征图大小不变,通道数由 64 增加至 256,稠密连接网络不仅可以减少模型参数,还可以利用稠密连接网络结构特点通过不断复用增强特征传播;stage2-4 中采用堆叠的轻量化倒残差结构来替换原网络中的 basicblock,通过引入轻量化的倒残差网络将模型参数和运算复杂度大大减少;最后在网络中增加改进的空间和通道双注意力机制,提高模型检测精度。实验表明,本文算法在 COCO 数据集上取得了较好效果。

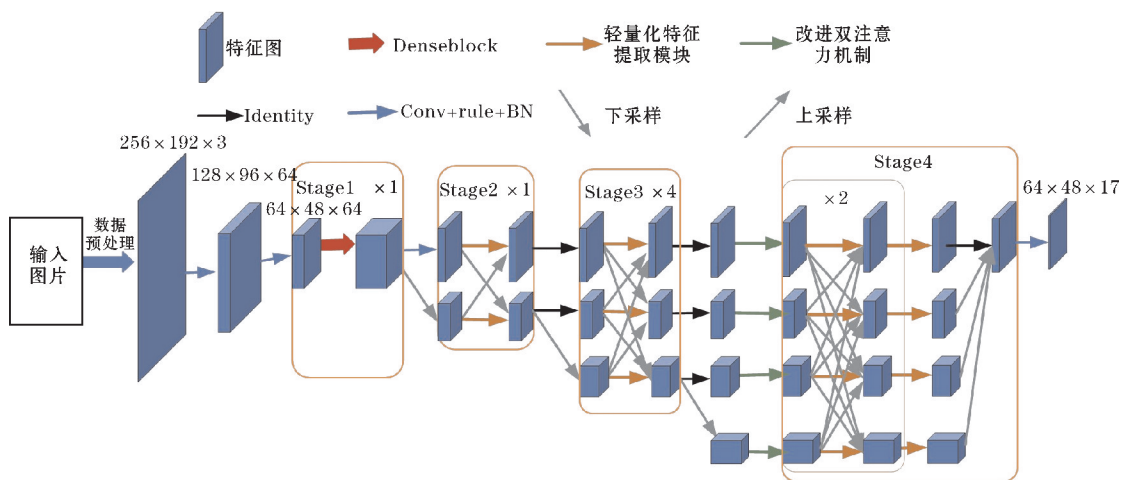


图1 ES-LHRNet 结构

1.1 稠密连接网络

经典残差网络^[9]中梯度可以通过恒等映射函数从后面的层流向前面的层,表示如下:

$$x_l = H_l(x_{l-1}) + x_{l-1}$$

这样的结构可能会阻碍网络中的信息流。稠密连接卷积网络^[10]采用让模块内的每一层接受它前面所用层的输出的方式缓解了梯度消失的问题,表示如下:

$$x_l = H_l([x_0, x_1, \dots, x_{l-1}])$$

相比于 ResNet, DenseNet 的特征复用能更好地提高层与层之间的信息流动,加强特征传播,大大减少参数数量。

参照 HRNet 中 stage1 特征提取过程中,保持原特征图大小不变,将通道数从 64 扩展到 256。设置 Denselayer 输出特征通道数为 32,每一次 Denselayer 的输入特征为第一次 Denselayer 的输入特征和之前每次的输出特征 Concat 的结果,在 Denseblock 结构中一次特征提取,过程如图 2 所示。且 Denseblock 由 6 次特征提取完成,每一次特征提取输入都融合前面的所

有输出结果,最后融合每一次特征提取结果得到通道数为 256 的特征作为最后的输出结果。图 3 为 Denseblock 中特征图的提取过程。

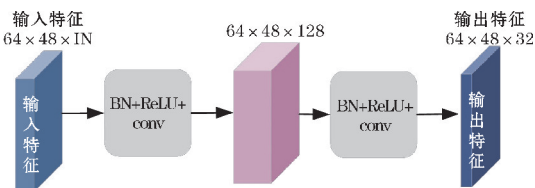


图2 Denselayer 特征提取过程

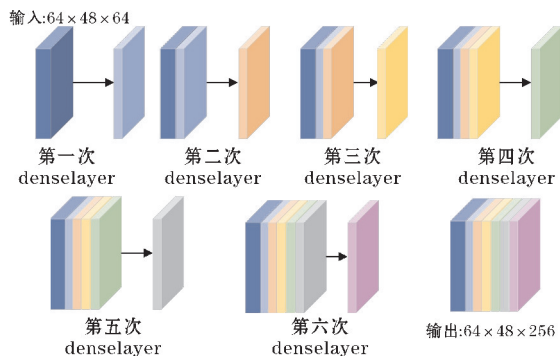


图3 Denseblock 特征提取过程

1.2 轻量化多尺度特征提取模块

经典残差网络的参数数量和计算量很大,往往给模型训练带来困难,在 ShuffleNetV2 中创新性地提出轻量级网络的评价标准,给轻量化大型网络提供了新的思路。对于人体姿态估计中关键点的多尺度特征提取,采用堆叠的轻量化倒残差特征提取模块替换原 HRNet 中的残差网络的 basicblock。通道划分将输入特征图平均分为两部分,一部分经过倒置残差提取特征处理后与直接与另一部分拼接。这样的特征提取过程保持了卷积层的输入特征矩阵与输出特征矩阵通道相等,有效控制了内存访问成本,同时通道划分为两个组也可以控制内存访问成本^[7]。轻量化的倒残差特征提取结构如图 4 所示,一个轻量化特征提取模块由 4 个轻量化倒残差结构堆叠而成。

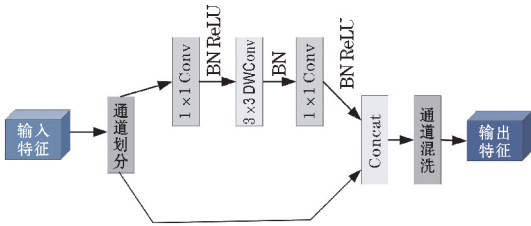


图 4 轻量化倒残差结构

1.3 双注意力机制

注意力机制具有模块化的特性,便于在网络中融合,在深度学习处理的各领域中有效地提升了网络的性能。通过注意力机制可以增加网络的表征力,让网络更加关注重要特征,抑制不必要特征。卷积运算是通过将跨通道信息和空间信息混合在一起提取特征的,Huang 等^[10]提出一种轻量化的通道注意力机制和空间注意力机制的双注意力机制。Woo 等^[11]提出一种不降维的局部跨通道交互策略,有效避免了降维对于通道注意力学习效果的影响。适当的跨通道交互可以在保持性能的同时显著降低模型的复杂性,达到通过少数参数的调整获得明显的效果增益^[12]。本文提出一种新的结合通道注意力机制和空间注意力机制的轻量化双注意力模块,在通道注意力中自适应 1d 卷积的输入和输出通道数都是 1,根据输入通道数自适应地计算卷积核的大小,输入通道数较大时卷积核也较大。图 5 为 ECA 通道注意力机制的处理过程。空间注意力不仅关注每个通道的比重,还关注每一个像素点的比重,图 6 为空间注意力机制处理过程。通过双注意力模块可以获取低分辨率特征图过程中的信息丢失问题以及增强高分辨率特征图的信息,图 7 为改进的双注意力机制。

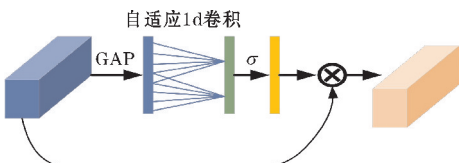


图 5 通道注意力机制

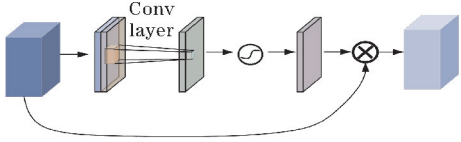


图 6 空间注意力机制



图 7 改进的双注意力机制

2 实验与分析

2.1 数据集简介

COCO2017 数据集中标注的人体检测关键点共 17 个,如表 1 所示。

表 1 COCO2017 数据集中人体骨骼关键点

编号	关键点名称	编号	关键点名称
0	鼻子	9	左手腕
1	左眼	10	右手腕
2	右眼	11	左臀
3	左耳	12	右臀
4	右耳	13	左膝
5	左肩	14	右膝
6	右肩	15	左脚踝
7	左肘	16	右脚踝
8	右肘		

2.2 评估标准

实验采用关键点相似度(object keypoint similarity, OKS)进行评估,验证标准平均正确率:

$$OKS = \frac{\sum_i \exp\left(-\frac{d_i^2}{2s^2k_i^2}\right) \delta(v_i > 0)}{\sum_i \delta(v_i > 0)}$$
$$AP = \frac{\sum_p \delta(OKS > i)}{\sum_p 1}$$

式中: d 为预测的关键点和 GT 之间的欧几里得距离。 s 为人体所占的面积大小的平方根,根据 GT 里的 box 计算得到。 k_i 第 i 个关键点的归一化因子,由数据集中所有 GT 计算的标准差得到,值越大,说明在整个数据集中对该点的标注效果越差;值越小,说明整个数据

集中对该点的标注效果越好。 $\delta_{(x)}$ 为 OKS 只计算>0 的所有关键点。 v_i 为第 i 个关键点的可见性。0 表示 GT 中不存在该点,1 表示 GT 中该点存在但是被遮挡,2 表示 GT 中该点存在且可见。 t 为设置的阈值,如果 $\text{OKS}>t$,那就说明检测成功,反之则说明不成功。

2.3 实验环境与设置

本文方法的实验环境为显卡 Nvidia GeForce GTX 2060,Windows 系统,python3.7,并使用 Pytorch 深度学习框架作为实验框架版本为 pytorch1.12.1 以及 cuda113。训练时通过随机图像旋转($[-45^\circ, 45^\circ]$),随机缩放($[0.65, 1.35]$)和随机水平翻转(0.5)将数据

集中的图像进行预处理。实际使用的输入图像大小为 256×192,训练周期设置为 210,GPU 的批量大小设置为 32,使用 Adam 优化器对网络进行优化,同时将初始学习率设置为0.001,多步长衰减,170 时衰减到 0.0001,200 时衰减到0.00001。

2.4 实验验证分析

将 ES-LHRNet 在 COCO2017 数据集上进行验证,同时与目前经典的姿态检测网络实验结果进行对比。表 2 和表 3 验证了本文所用方法可以在参数量较少的情况下提升网络的检测准确率。

表 2 COCO2017 数据集下与其他大型网络的验证比较

model	骨干网络	预训练	params	GFLOPs	AP	AP ⁵⁰	AP ⁷⁵	AP ^M	AP ^L	AR
8-stage Hourglass	8-stage Hourglass	N	25.1M	14.3	66.9	—	—	—	—	—
CPN	ResNet-50	Y	27.0M	6.20	68.6	—	—	—	—	—
SimpleBaseline	ResNet-50	Y	34.0M	8.90	70.4	88.6	78.3	67.1	77.2	76.3
HRNetV1-32	HRNetV1-W32	N	28.5M	7.10	73.4	89.5	80.7	70.2	80.1	78.9
MSANet ^[13]	HRNetV1-W32	N	28.1M	6.90	77.6	94.6	84.8	73.9	84.0	81.1
文献[14]	HRNetV1-W32	N	17.5M	4.6	72.3	88.1	79.5	68.7	79.1	77.9
ours	HRNetV1-W32	N	3.96M	2.16	67.7	90.5	75.9	65.0	71.8	71.0

表 3 COCO2017 数据集下与其他轻量型网络的验证比较

model	骨干网络	预训练	params	GFLOPs	AP	AP ⁵⁰	AP ⁷⁵	AP ^M	AP ^L	AR
MobileNetV2	MobileNetV2	Y	9.6M	1.48	64.6	87.4	72.3	61.1	71.2	70.7
ShuffleNetV2	ShuffleNetV2	Y	7.6M	1.28	59.5	85.4	66.3	56.6	66.2	66.4
Small HRNet	Small HRNet	N	1.3M	0.54	55.2	83.7	62.4	52.3	61.0	62.1
Lite-HRNet-18	Lite-HRNet-18	N	1.1M	0.20	64.8	86.7	73.0	62.1	70.5	71.2
Lite-HRNet-30	Lite-HRNet-30	N	1.8M	0.20	67.2	88.0	75.0	64.3	73.1	73.3
ours	ours	N	3.96M	2.16	67.7	90.5	75.9	65.0	71.8	71.0

2.5 消融实验

为验证本文改进方法对算法提升的效果,在 CO-

CO2017 数据集上进行消融实验。消融实验结果如表 4 所示,表明本文改进方法提升了轻量型高分辨率关键点检测算法的准确率。

表 4 本文方法的消融实验

实验	轻量化 HRNet	Denseblock	注意力	params	GFLOPs	AP	AP ⁵⁰	AP ⁷⁵	AP ^M	AP ^L	AR
实验 1	✓	×	×	3.88 M	1.90	67.1	89.5	74.7	64.5	71.4	70.6
实验 2	✓	✓	×	3.96 M	2.16	66.1	89.3	73.6	63.2	70.4	69.5
实验 3	✓	×	✓	3.88 M	1.90	67.5	90.3	75.0	65.0	71.8	71.0
实验 4	✓	✓	✓	3.96 M	2.16	67.7	90.5	75.9	65.0	71.8	71.0

图 8 为模型训练过程中以及后 60 个 epoch 的 mAP 和 AP⁵⁰值的变化过程,从图 8 得到:通过实验 1 发现仅经过倒残差轻量化的方法在模型训练中收敛过程出现了精度骤降。通过实验 2 发现增加 DenseBlock 的轻量化网络改善了精度骤降的问题,这可能是稠密连接网络对特征的不断复用得到的好结果,但是相比

倒残差轻量化的方法的精度普遍下降。这可能是因为在瓶颈处使用堆叠的倒残差结构将保持特征的通道数不变。实验 1 中加了一个普通卷积将特征从 64 层处理为 256 层,但仅从这两个实验不能确定稠密网络结构对本文方法的作用。通过实验 3 和实验 4 可以看出引入 ESAM 注意力机制的两个方法精度明显高于没

有引入注意力机制的实验方法,说明本文 ESAM 注意力机制在人体关键的检测中的有效性。同时发现实验 4 中增加 DenseBlock 的方法精度高于实验 3 的方法, DenseBlock 和 ESAM 同时使用对本文轻量型算法精度

提升有效。图 9 为算法训练过程中损失值的收敛情况(第一个 epoch 中 loss 值较高没有在图中表示出来),可以发现 ES-LHRNet 更快收敛,说明本文改进方法的可靠性。检测结果可视化如图 10 所示。

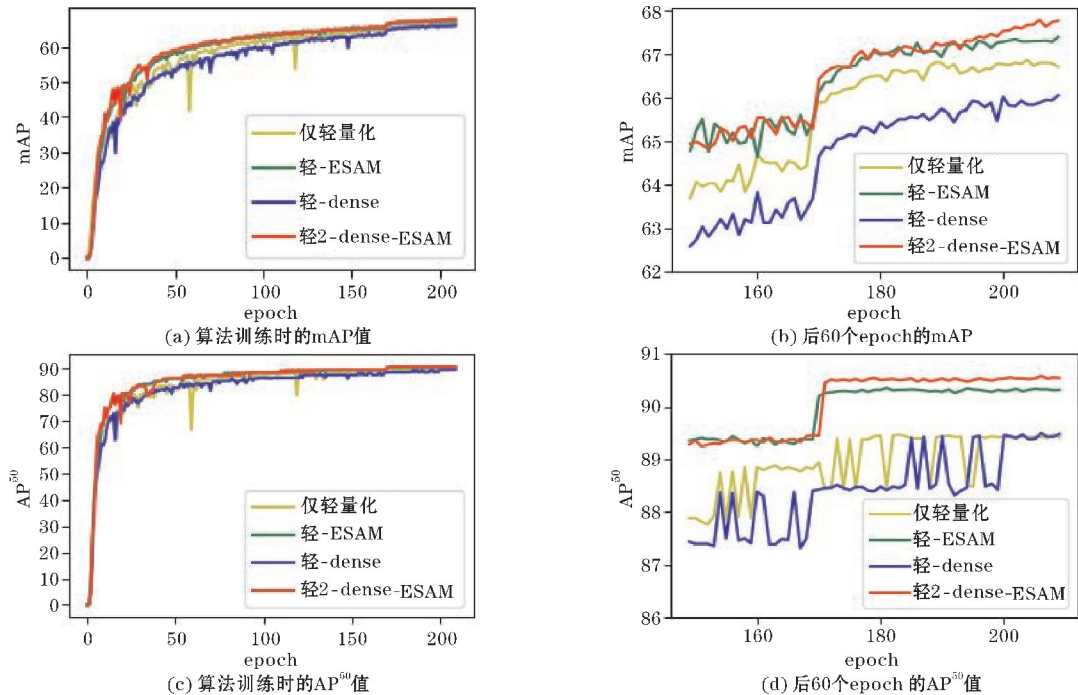


图 8 算法训练及后 60 个 epoch 的 mAP 和 AP⁵⁰ 值

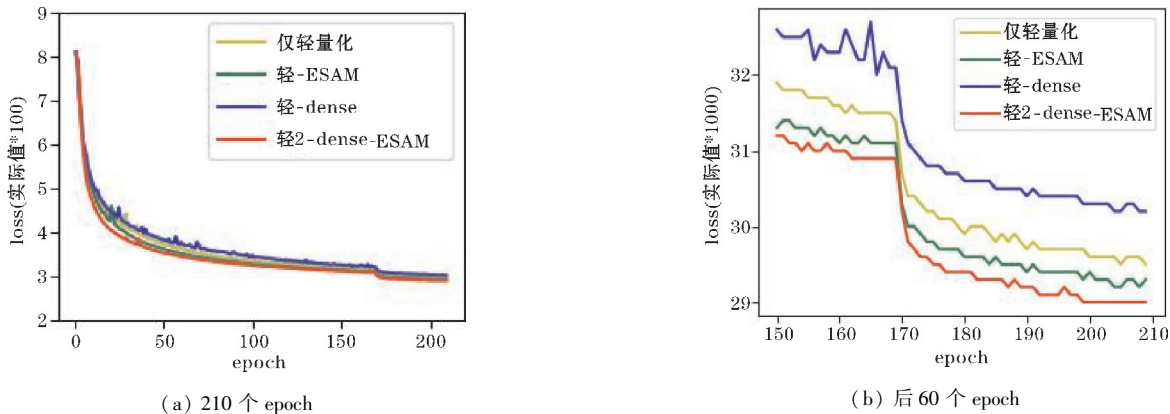


图 9 训练时 loss 曲线

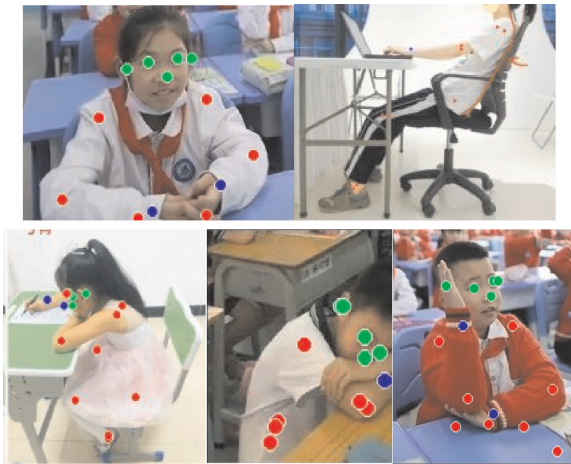


图 10 检测结果可视化

3 结束语

为解决大型姿态检测算法难训练,轻量型网络准确率低,本文以 HRNet 网络为研究基础提出一种较高准确率的轻量型人体关键点检测算法。结果表明,在输入图像预处理为 256×192 大小的情况中,相比原 HRNet、ES-LHRNet 的 AP⁵⁰ 结果提升了 1%,相比于轻量型 Lite-HRNet-30,本模型的 AP⁵⁰ 结果提升了 0.5%,表明该网络提升了人体关键点的准确度。消融实验结果表明,融合稠密网络和改进双注意力机制在关键点检测中的有效性。姿态检测研究可以让计算机具有感知人的行为的能力,目前模型训练所需要的时间成本

仍然较大,未来的研究应该在兼顾模型准确率和参数量大小的同时提升模型的训练速度。

参考文献:

- [1] Toshev A, Szegedy C. Deeppose: Human pose estimation via deep neural networks [C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2014:1653–1660.
- [2] Wei S E, Ramakrishna V, Kanade T, et al. Convolutional pose machines [C]. Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2016:4724–4732.
- [3] Cao Z, Simon T, Wei S E, et al. Realtime multi-person 2d pose estimation using part affinity fields [C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017:7291–7299.
- [4] Chen Y, Wang Z, Peng Y, et al. Cascaded pyramid network for multi-person pose estimation [C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2018:7103–7112.
- [5] Sun K, Xiao B, Liu D, et al. Deep high-resolution representation learning for human pose estimation [C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019:5693–5703.
- [6] Yu C, Xiao B, Gao C, et al. Lite-hrnet: A lightweight high-resolution network [C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021:10440–10450.
- [7] Ma N, Zhang X, Zheng H T, et al. Shufflenet v2: Practical guidelines for efficient cnn architecture design [C]. Proceedings of the European conference on computer vision (ECCV). 2018:116–131.
- [8] 卢官明, 卢峻禾, 陈晨. 基于深度学习的二维人体姿态估计研究进展 [J]. 南京邮电大学学报 (自然科学版), 2024, 44(1): 44–55.
- [9] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition [C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016:770–778.
- [10] Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks [C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017:4700–4708.
- [11] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module [C]. Proceedings of the European conference on computer vision (ECCV), 2018:3–19.
- [12] Wang Q, Wu B, Zhu P, et al. ECA-Net: Efficient channel attention for deep convolutional neural networks [C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020:11534–11542.
- [13] 李丽, 张荣芬, 刘宇红, 等. 基于多尺度注意力机制的高分辨率网络人体姿态估计 [J]. 计算机应用研究, 2022, 39(11): 3487–3491.
- [14] 朱翠涛, 李博. 基于高分辨率网络的人体姿态估计 [J]. 中南民族大学学报 (自然科学版), 2023, 42(2): 229–237.

Lightweight Human Pose Estimation Algorithm based on Improved Dual Attention Mechanism

TANG Zhixuan^{1,2}, GAO Yuxiang^{1,2}

(1. College of Communication Engineering, Chengdu University of Information Technology, Chengdu 610225, China; 2. Meteorological Information and Signal Processing Key Laboratory of Sichuan Education Institutes, Chengdu 610225, China)

Abstract: To improve the accuracy of the lightweight human posture detection algorithm, a new dual-attention mechanism of lightweight high-resolution human pose estimation network ES-LHRNet was proposed. Based on the overall framework of HRNet (high-resolution network) as a reference, this paper carried out feature extraction by Densenet and stacked lightweight inverse residual structure, and a new dual attention mechanism that integrates spatial attention mechanism and channel attention mechanism to capture location information and channel information promotion algorithm. Compared with HRNet, the number of parameters in this model is reduced by 86%, the computational complexity is reduced by 73%, and the proposed ESAM improves the map of detection results on dataset MS COCO 2017 by 1.6 percentage.

Keywords: human pose estimation; high-resolution network; attention mechanism; feature reuse; lightweight network