ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 40,No. 6,Nov. 2025,pp. 1625－1636
DOI:10. 16337/j. 1004-9037. 2025. 06. 019

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于局部通道编码的轻量化人体姿态估计算法

徐新智，何　宏

(上海理工大学健康科学与工程学院,上海 200093)

摘　要：针对当前姿态估计模型存在计算复杂度高以及参数量大的问题,本文提出一种轻量级姿态估计算法。首先,在特征提取过程中引入局部通道编码(Partial channel encoding，PCE)模块,结合卷积神经网络与视觉编码器的优势,分别提取图像的局部特征和全局特征;接着在多尺度特征融合过程中引入加权特征融合,增强模型的多尺度特征融合能力以避免模型轻量化带来的精度降低的问题;之后在回归预测的过程中将人体检测和分类部分共享检测头,提高模型在姿态估计任务中的识别效率;最后将CIoU损失函数更换为PIoU损失函数,让模型更注重对中高质量检测框的识别准确度。实验结果表明,本文提出的模型相比于基础模型,参数量下降27%,计算量下降18%,准确度提升0.2%,既保证了识别的准确度,又可以实现检测算法的轻量化,为实现实时准确的姿态估计提供了有效手段。

关键词：姿态估计;编码器;轻量化;共享权重;多尺度融合

中图分类号：R771.3; R581; TP391.4　　**文献标志码：**A

Lightweight Human Pose Estimation Algorithm Based on Partial Channel Encoding

XU Xinzhi，HE Hong

(School of Health Science and Engineering，University of Shanghai for Science and Technology，Shanghai 200093，China)

Abstract：Aiming at the problems of high computational complexity and large number of parameters in the current pose estimation model，this paper proposes a lightweight pose estimation algorithm. Firstly，the partial channel encoding (PCE) module is introduced in the feature extraction process，and the local and global features of the image are extracted respectively by combining the advantages of convolutional neural network and visual encoder. Then，the weighted feature fusion is introduced in the process of multi-scale feature fusion to enhance the multi-scale feature fusion ability of the model and avoid the problem of reduced accuracy caused by model lightweight. Then，in the process of regression prediction，the detection head of the human detection and classification parts is shared to improve the recognition efficiency of the model in the pose estimation task. Experimental results show that compared with the basic model，the proposed model reduces the number of parameters by 27% and the amount of computation by 18%，and increases the accuracy by 0.2%. It not only ensures the accuracy of recognition，but also realizes the lightweight of the detection algorithm，providing an effective means to achieve real-time accurate pose estimation.

Key words: pose estimation; encoder; lightweight; shared weight; multi-scale fusion

---

基金项目：国家科技部项目(G2021013008);教育部中国高校产学研创新基金(2023RY011);上海理工大学医工交叉重点项目(1020308405,1022308502)。
收稿日期：2024-10-08;**修订日期：**2025-02-20

引　言

随着人工智能在机器视觉领域的发展，人体姿态估计是机器视觉方向的重要挑战之一，它在许多领域都有广泛的应用前景，包括虚拟现实、运动和健康检测、智能监测等。传统方法利用图结构和形变部件模型设计2D人体部件检测器，但过度依赖人工特征提取，数据获取难度大，因此众多研究人员将基于深度学习的目标检测算法应用到人体姿态估计中。

目前，基于深度学习的多人姿态估计可分为自下而上和自上而下两种结构。自下而上过程是先提取出图像中所有人的人体关节点，再使用算法完成单人关节点分组任务，实现多人姿态估计。例如Pishchulin等[1]提出DeepCut算法，使用Faster R-CNN检测人身体部件，后使用整数线性规划算法匹配人体部件实现人体姿态估计，后续又设计了DeeperCuter使用更深的ResNet检测人体部件，并设计使用了图像条件化的匹配项，提升人体部件匹配效率。Cao等[2]提出将身体部位彼此关联以形成完整的人体姿态的表示方法，称为部件关联场(Part affinity field, PAF)，该算法包含两路网络分支分别预测行人关键点和部件关联场，通过贪婪推理法解析PAF预测出人体的关键点。Kreiss等[3]在PAF基础上提出用来定位身体部位的部件强度场(Part intensity field, PIF)，将二者混合形成复合场，使得该算法在低分辨率图像上优于以前的算法。自上而下是先进行人体检测，提取出单人图像后输入到姿态估计模型中提取关键点。例如，He等[4]提出了Mask R-CNN，该算法在行人检测器Faster R-CNN的基础上，增加了用于单人姿态估计的网络分支，通过这种联合优化方法，可以对行人检测任务和单人姿态估计任务进行统一的处理和优化误差。Fang等[5]提出多人姿态估计模型AlphaPose，该模型使用YOLOv3作为人体检测器，并在单人姿态估计部分提出一种对称变换空间网络(Symmetric spatial transformer network, SSTN)，可以从不准确的边界框中提取高质量的单人区域。但是这些分两个阶段的姿态估计方法存在一些问题，比如人体拥挤或遮挡时，预测准确率降低，其次分离模型需要分开训练，无法端到端优化等。针对上述问题，Debapriya等[6]提出了一种多人姿态估计算法YOLO-Pose，使用无热力图联合检测方法实现了端到端的训练并且优化目标关键点。然而，更好的性能往往需要消耗更多算力和存储空间，这对于计算能力和存储空间有限的设备来说并非易事。

针对上述问题，本文提出一种基于YOLOv8Pose的轻量化姿态估计算法，在特征提取层中使用混合编码器结构，提升模型对局部和全局的特征提取能力，同时引入加权双向特征金字塔网络(Bi-directional feature pyramid network, BiFPN)提升模型的多尺度特征融合能力，并在关键点和检测框回归中提出共享卷积减少参数量，最后使用强聚焦损失函数PIoU提升模型对人体的识别能力。模型既可以保证识别的准确度，又可以实现检测算法的轻量化，为实现实时准确的姿态估计提供了有效手段。

1　相关工作

1.1　注意力机制

注意力机制在自然语言处理中取得了巨大成功后，逐渐运用于图像处理邻域，注意力机制通过对输入数据中不同部分的重要性进行动态加权，使模型能够集中注意力在最有用的信息上，从而提升模型在复杂任务中的表现和解释能力。Hu等[7]设计了压缩和扩张(Squeeze-and-excitation, SE)模块，通过对通道中的有效和无效特征信息进行增强和抑制处理以获取卷积特征通道之间的关联性，进而生成高质量的特征图。Woo等[8]提出卷积注意力模块(Convolutional block attention module, CBAM)，先后使用通道注意力和空间注意力，经过这两个注意力模块的串行操作，最初的特征图就经过了通道和空间两个注意力机制的处理，自适应细化特征。张鹏等[9]提出深度非对称瓶颈(Depth-wise asymmetric bottleneck, DAB)模块，降低边缘效应对实时语义分割效果的影响，又设计了注意力融合模块自上而下地融合不同尺度下

的特征信息,实现全局特征信息下的有效交互,增强了网络对重要特征的表达。Dosovitskiy等[10]提出图像Transformer(Vision Transformer, ViT),将图像划分成中等大小的图像块并转换一系列固定长度的补丁嵌入,并通过Transformer架构进行图像分类,从而实现了速度和精度的均衡。

1.2 轻量级神经网络

轻量级神经网络可以在有限的计算资源里实现良好性能,适用于嵌入式设备及对时延要求严苛的实时应用。Howard等[11]将视觉几何组(Visual geometry group, VGG)中的标准卷积层换成深度可分离卷积提出了MobileNetV1,该网络在准确率小幅度降低的前提下大幅度减少模型参数与运算量。Lightweight OpenPose[12]针对OpenPose计算成本较高、综合性能较差等问题,使用MobileNet替代OpenPose骨干网络,其在多人姿态估计领域中表现出了良好的检测精度。Tan等[13]提出EfficientNet,采用固定的比例分别缩放网络的宽度、深度以及输入图像的分辨率以提高网络性能。当然使用固定的比例也存在缺陷,当训练图像的尺寸很大时,会导致训练速度非常慢,并且在网络浅层中使用深度卷积训练速度也会很慢。Li等[14]提出一种Dite-HRNet模型,利用分离卷积、逐点卷积和通道混洗算子设计了动态多尺度上下文(Dynamic multi-scale context, DMC)和动态全局上下文(Dynamic global context, DGC)两个高效的卷积模块来替换Small HRNet中的残差结构,在小幅降低模型运算复杂度的同时提高了精度。贾子豪等[15]提出轻量型三角式卷积层改进YOLOv5检测头并在特征融合部分引入CBAM,最终实现模型的轻量化。

2 模型结构

本文提出的基于局部通道编码的轻量级姿态估计模型PCE_LitePose(Partial channel encoding_LitePose)以YOLOv8Pose作为基础网络框架,可以同时检测出人体的关键点和人体检测框。该网络分3部分:Backbone负责多尺度特征提取,Neck负责多尺度特征融合以及Head负责人体关键点以及检测框回归。其中,在Backbone中引入局部通道编码(Partial channel encoding, PCE)结构,在保证轻量化的同时提升模型对全局特征的理解能力;在Neck中引入加权特征融合BiFPN,通过对每个输入特征赋予独特权重,使网络能动态调整权重以判断输入特征的重要性;基础模型在人体检测、分类以及姿态估计中都使用相同结构,但是人体检测和分类任务相比于姿态估计难度较低,这就造成模型参数过剩。因此,本文提出局部权重共享头(Partial share weight header, PSWH),该模块使用共享卷积进行人体检测和分类任务,模型结构如图1所示。

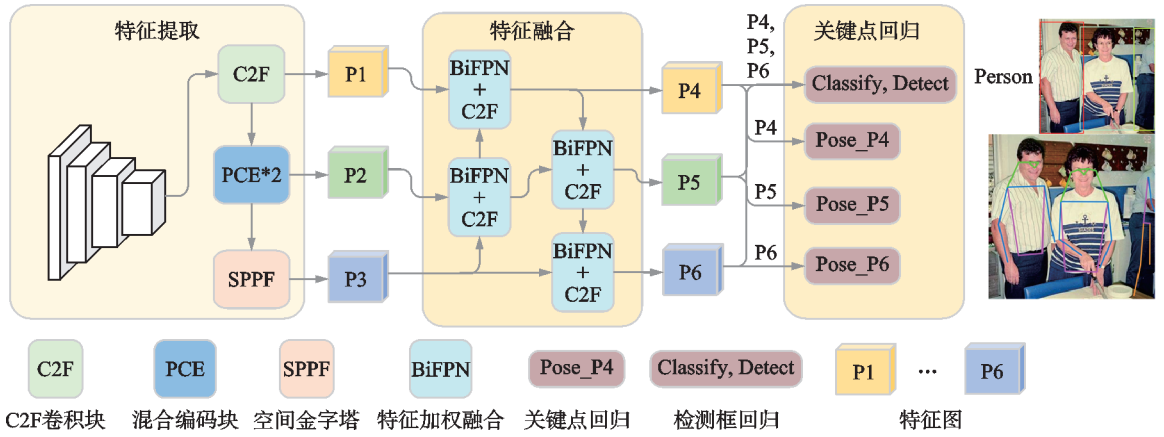

图1 PCE_LitePose模型结构

Fig.1 PCE_LitePose model structure

2.1 局部编码结构PCE

为了实现更准确的人体关键点回归，特征图的位置关联与全局特征信息的利用尤为重要。绝大部分卷积神经网络模型都通过堆叠卷积层数或提升卷积核大小的方法来获取图像全局特征信息，但是重复堆叠卷积和增加卷积核大小都会大大提升模型的参数量。首先，视觉编码器ViT架构提供了一种创新的解决方案，通过多头自注意力机制达到捕捉长距离依赖关系和全局上下文信息的目的，对物体的位置变化更加鲁棒，可适应不同尺度和旋转下的人体姿态；其次，在使用卷积神经网络的过程中，不同通道的特征存在高度的相似性，对部分通道进行常规卷积，剩余部分通道的特性保持不变，降低了计算复杂度，也实现了快速高效的神经网络。所以本文提出一种轻量级的特征提取模块PCE，通过对部分通道使用卷积神经网络提取图像的局部特征，对其余通道使用视觉编码器ViT结构并结合卷积门控线性单元(Convolutional gated linear unit, CGLU)提取图像的全局特征，模块计算流程如图2所示。

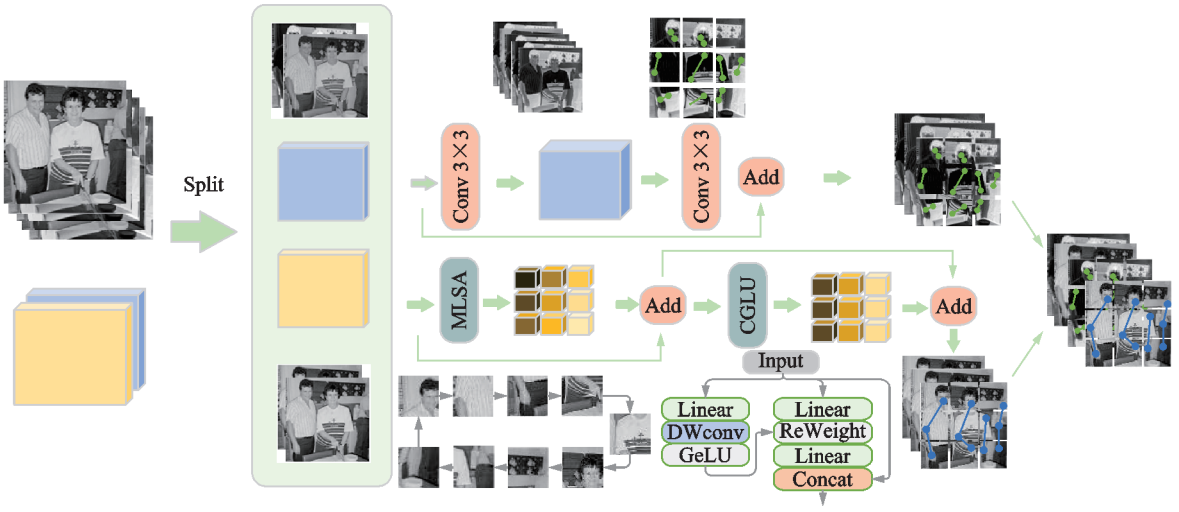

图2 PCE模块计算流程

Fig.2 PCE module process

假设输入尺寸为$F\in R^{H\times W\times C}$，对于部分通道特征图$R^{H\times W\times C_e}$，使用卷积神经网络+残差结构来捕捉局部特征$F_e$，其计算公式为

$$F_e=\mathrm{Conv}\left(\mathrm{Conv}\left(R^{H\times W\times C_e}\right)\right)+R^{H\times W\times C_e} \tag{1}$$

对于剩余通道特征图$R^{H\times W\times C_t}$，使用视觉编码器中多头自注意力+通道混合器结构捕捉全局特征F_t，其计算公式为

$$R^{N\times(P^2\times C_t)}=R^{N\times D}=R^{H\times W\times C_t} \tag{2}$$

$$Z=[x_{\mathrm{class}};x_{\mathrm{p}}^1E;x_{\mathrm{p}}^2E,\cdots;x_{\mathrm{p}}^NE]+E_{\mathrm{pos}}\quad E\in R^{N\times(P^2\times C_t)},E_{\mathrm{pos}}\in\mathrm{Patch}\left(R^{D\times(N+1)}\right) \tag{3}$$

$$Z_l'=\mathrm{MLSA}(Z)+Z \tag{4}$$

$$Z_l=\mathrm{CGLU}(Z_l')+Z_l' \tag{5}$$

式中：N代表序列长度，$P^2\times C_t$表示序列中图像块的维度，D表示序列中图像块的维度，Z表示包含位置编码的图像块，Z_l'表示包含多头自注意力权重的特征图，Z_l表示经过通道混合处理后的特征图，MLSA(•)表示多头自注意力计算，其计算公式为

$$V=W^vZ,Q=W^qZ,K=W^kZ \tag{6}$$

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{7}$$

$$\text{head}_i=\text{Attention}(\boldsymbol{Q}\boldsymbol{W}_i^q,\boldsymbol{K}\boldsymbol{W}_i^k,\boldsymbol{V}\boldsymbol{W}_i^v) \tag{8}$$

$$\text{MLSA}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{Concat}(\text{head}_1,\text{head}_2,\cdots,\text{head}_h)\boldsymbol{W}^O \tag{9}$$

式中:$\boldsymbol{W}^q$、$\boldsymbol{W}^k$和$\boldsymbol{W}^v$分别表示$\boldsymbol{Q}$、$\boldsymbol{K}$和$\boldsymbol{V}$的线性变换矩阵,CGLU表示通道混合器,其中一个是通过激活函数激活的元素乘法,每个Token的gating信号都来自Token本身,并且其感受野大小不超过value分支的感受野,在GLU的gating分支的激活函数之前,简单地添加一个最小形式的卷积,可以使它的结构符合基于最近邻特征的门控通道注意力的设计概念,并将其转换为门控通道注意力机制。

2.2 加权特征融合BiFPN

经过卷积神经网络提取后的特征P1与PCE模块特征提取后的特征P2,P3的信息不同。前一步将通道分开处理不利于通道间的特征融合,而PAN结构将不同尺寸的特征直接通道拼接,忽视了不同尺寸的特征贡献度不同的问题。加权特征融合BiFPN[16]是在PAN结构的基础上优化出来的,采纳双向融合的理念,对自顶向下和自底向上的信息流动路径进行了重新规划,并且对每个需要融合的特征赋予独特权重,使网络能动态调整权重以判断输入特征的重要性。基础网络中采用的上采样模块为UpSample模块。该模块主要是采用最近邻插值的方法进行上采样,没有利用到特征图的语义信息,而且感知域通常都很小,而CARAFE模块[17]通过上采样核预测和特征重组可以完成上采样,支持具体内容具体处理的操作,具有自适应内核。BiFPN和CARAFE的结构如图3所示。

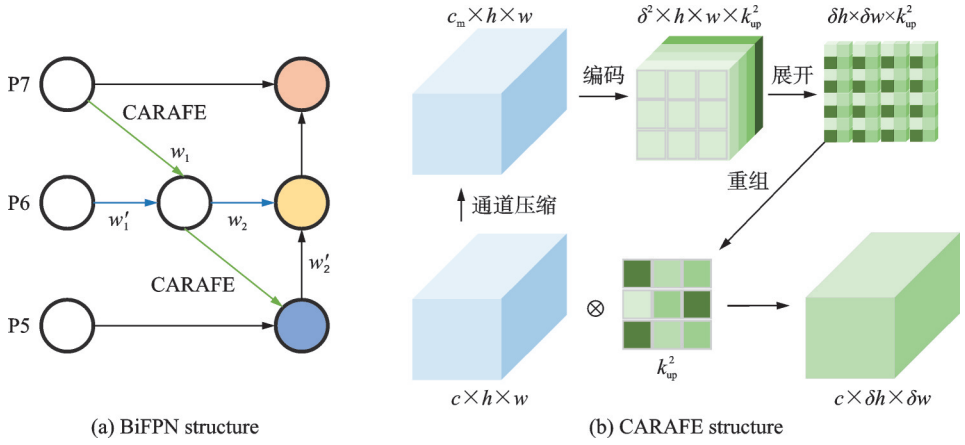

图3 BiFPN和CARAFE结构

Fig.3 Structures of BiFPN and CARAFE

BiFPN的计算公式为

$$P_7^{\text{out}}=\text{Conv}(P_7^{\text{in}}) \tag{10}$$

$$P_6^{\text{mid}}=\text{Conv}\left(\frac{w_1\cdot P_6^{\text{in}}+w_1'\cdot\text{CARAFE}(P_7^{\text{in}})}{w_1+w_1'+\epsilon}\right) \tag{11}$$

$$P_6^{\text{out}}=\text{Conv}\left(\frac{w_2\cdot P_6^{\text{mid}}+w_2'\cdot\text{CARAFE}(P_5^{\text{out}})}{w_2+w_2'+\epsilon}\right) \tag{12}$$

式中:P_6^{out}表示P6层的输出特征,P_6^{mid}表示P6层的中间特征,P_6^{in}表示P6层的输入特征,Conv表示卷积

操作,CARAFE表示上采样操作。CARAFE的计算流程如下:(1)对输入特征图进行通道压缩,输入尺寸为$h\times w\times c$,压缩处理后尺寸为$h\times w\times c_{\mathrm{m}}$,大大减小了计算量;(2)将通道压缩后的特征图进行上采样核预测,利用大小为$k_{\mathrm{up}}\times k_{\mathrm{up}}$的卷积核对上述压缩特征图进行内容编码,得到尺寸为$\delta^2\times k_{\mathrm{up}}^2\times h\times w$的特征图;(3)该特征图在通道维度上展开后尺寸变为$k_{\mathrm{up}}^2\times\delta h\times\delta w$;(4)经过归一化和重组后与输入特征图进行卷积得到上采样结果。

2.3 局部权重共享头PSWH

为了获取多人的人体检测和关键点,需要对不同尺度的人体目标进行预测,但是人体检测和分类任务相比于姿态估计难度较低,对人体检测和姿态估计使用相同结构的检测头降低了模型预测效率。因此,本文提出局部权重共享头PSWH,该模块使用共享卷积进行人体检测和分类任务,共享卷积可以降低存储需求和计算成本,但限制了模型在不同尺度特征独立学习参数的能力,所以在使用共享卷积前使用逐点卷积,用于调整特征图的通道数量并引入非线性变换。基础Head模块和PSWH模块结构如图4所示。

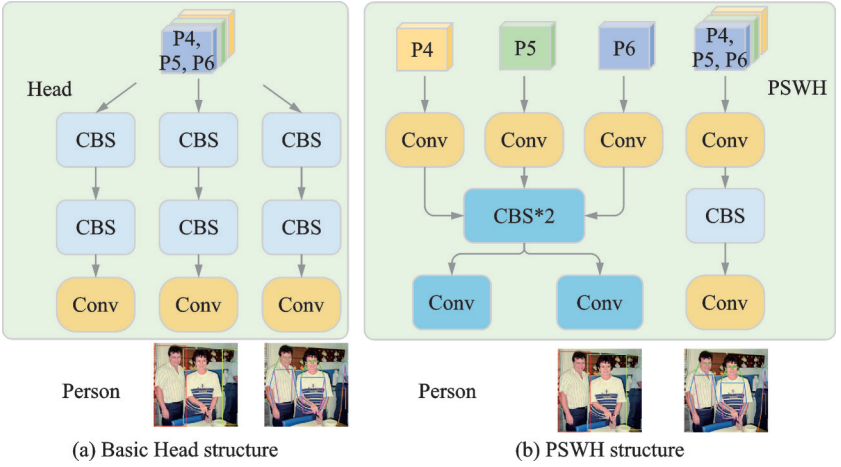

图4 基础Head和PSWH结构

Fig.4 Structures of basic Head and PSWH

一个Head的计算复杂度为

$$\mathrm{FLOPs}_{\mathrm{conv}}=(2\times c_0\times k^2-1)\times h\times w\times c_1 \tag{13}$$

$$\mathrm{FLOPs}_{\mathrm{Head}}=6\times\mathrm{FLOPs}_{\mathrm{conv3\times3}}+3\times\mathrm{FLOPs}_{\mathrm{conv1\times1}}=h\times w\times c_1\times[6\times(18c_0-1)+3c_0] \tag{14}$$

式中:$\mathrm{FLOPs}_{\mathrm{conv}}$表示一个卷积的计算复杂度,$\mathrm{FLOPs}_{\mathrm{Head}}$表示一个尺度的Head计算复杂度。3个Head的计算复杂度为

$$3\mathrm{FLOPs}_{\mathrm{Head}}=3\times h\times w\times c_1\times[6\times(18c_0-1)+3c_0]=$$
$$33\times h\times w\times c_1\times c_0-18\times h\times w\times c_1 \tag{15}$$

式中:h和w分别表示特征图的长度和宽度,k表示卷积核的大小,c_0和c_1分别表示输入通道数和输出通道数。在检测过程中有通道数不同的特征图从Neck部分输入到Head中,其中P4为256×80×80,P5为512×40×40,P6为256×20×20。面对通道数不同的特征无法使用相同的卷积层完成特征提取,在共享卷积前使用逐点卷积进行通道数统一,后通过共享CBS模块完成特征提取。因为目标检测和分类任务通道数不同,再使用一次共享逐点卷积完成不同的任务需求。考虑到姿态估计比人体检测更具挑

战性,因为它不仅要求精确定位人体的关键点,还需要处理复杂的特征学习、遮挡问题、长距离依赖关系以及多样化的姿态和角度,难度较高,所以在对于Neck部分每个尺度的特征图使用非共享的CBS模块和逐点卷积进行姿态估计,PSWH模块模块的计算复杂度为

$$\mathrm{FLOPs}_{\mathrm{PSWH}}=[11\times c_0+2\times(18c_0-1)]\times h\times w\times c_1=(47c_0-2)\times h\times w\times c_1 \tag{16}$$

在$\mathrm{FLOPs}_{\mathrm{PSWH}}$和$3\mathrm{FLOPs}_{\mathrm{Head}}$中$h$、$w$、$c_0$、$c_1$值相同,并且$c_0>c_1$,显然使用共享卷积后的计算量$\mathrm{FLOPs}_{\mathrm{PSWH}}$比非共享的$3\mathrm{FLOPs}_{\mathrm{Head}}$计算量更低。

2.4 强聚焦损失函数PIoU

PCE_LitePose模型的损失函数主要分两部分:人体关键点损失和目标检测置信度损失。针对人体关键点损失,采用目标关键点相似度(Object keypoint similarity, OKS)为模型评价指标,对于特定部位的关键点倾斜重要性,比如在耳朵、鼻子、眼睛会比肩膀、膝盖、臀部等在像素级别上受到更多的错误惩罚;针对每一个单独的关键点计算OKS指标,并累加到最终的OKS损失,公式为

$$L_{\mathrm{kpts}}(s,i,j,k)=1-\sum_{n=1}^{N_{\mathrm{kpts}}}\mathrm{OKS}=1-\frac{\sum_{n=1}^{N_{\mathrm{kpts}}}\exp\left(\frac{d_n^2}{2s^2k_n^2}\right)\delta(v_n>0)}{\sum_{n=1}^{N_{\mathrm{kpts}}}\delta(v_n>0)} \tag{17}$$

式中:n表示关键点类型;d_n表示单个关键点的预测值与实际值的欧式距离;k_n是一个归一化因子,是通过对所有样本的人工标注和真实值的统计标准差;v_n表示当前关键点是否可见;δ用于将可见点选出来进行计算的函数;s是一个尺度因子,数值为人体检测框的面积平方根。

目标检测损失包含DFL损失[18]与CIoU损失[19],用于优化模型对人体预测框的置信度,基于目标中心点计算每个边框的DFL损失值,后基于预测框与实际框计算IoU损失值。其中,DFL以交叉熵的形式去优化与标签最接近的左右2个位置的概率,从而让网络更快地聚焦到目标位置及邻近区域的分布。计算公式为

$$\mathrm{DFL}(\mathcal{S}_i,\mathcal{S}_{i+1})=-((y_{i+1}-y)\ln\mathcal{S}_i+(y-y_i)\ln\mathcal{S}_{i+1}) \tag{18}$$

式中:$\mathcal{S}_i$和$\mathcal{S}_{i+1}$分别为模型的输出预测值和临近预测值;y为标签值;y_i和y_{i+1}为最接近标签的两个整数值。

在训练过程中,CIoU损失函数理论上会将锚框与目标框之间的位置和尺寸匹配,然而,在实际过程中,CIoU的回归过程有时会首先扩大锚框的大小以尝试增加与目标框的重叠程度,这导致了预测过程变得复杂且效率低下,需要更多的训练轮次才能收敛,并且CIoU中的惩罚项设计未能充分考虑目标框的实际尺寸,其构成在某些情况下不能准确地反映锚框与目标框的差异。CIoU计算公式为

$$\mathrm{CIoU}=\mathrm{IoU}-\left(\frac{\rho^2(b,b^{\mathrm{gt}})}{c^2}+\frac{v^2}{(1-\mathrm{IoU})+v}\right) \tag{19}$$

$$v=\frac{4}{\pi^2}\left(\arctan\frac{w^{\mathrm{gt}}}{h^{\mathrm{gt}}}-\arctan\frac{w}{h}\right)^2 \tag{20}$$

$$\mathrm{Loss}_{\mathrm{CIoU}}=1-\mathrm{IoU}+\frac{\rho^2(b,b^{\mathrm{gt}})}{c^2}+\alpha v \tag{21}$$

式中:IoU表示预测框和实际框的交并比;b和b^{gt}分别表示预测框和实际框的中心点;ρ是用来计算两个中心点变量的欧氏距离;c代表预测框和实际框所形成的外包最小矩形的对角线长度;w和h分别表示预测边界框的宽度和高度;w^{gt}和h^{gt}分别表示真实边界框的宽度和高度;α为权重函数。

PIoU[20]作为损失函数比CIoU更快地收敛，在前一步中对不同尺度的特征图使用共享卷积层削弱了模型对人体锚框的能力，所以在PIoU中引入了1个注意层，通过1个超参数让模型更聚焦于中等和高质量锚框的能力，提高了目标检测器的性能，其公式为

$$P=\frac{\left(\frac{\mathrm{d}w_1}{w_{\mathrm{gt}}}+\frac{\mathrm{d}w_2}{w_{\mathrm{gt}}}+\frac{\mathrm{d}h_1}{h_{\mathrm{gt}}}+\frac{\mathrm{d}h_2}{h_{\mathrm{gt}}}\right)}{4} \tag{22}$$

$$\mathrm{PIoU}=\mathrm{IoU}+\mathrm{e}^{-p^2}-1 \quad -1\leqslant\mathrm{PIoU}\leqslant 1 \tag{23}$$

$$L_{\mathrm{PIoU}}=1-\mathrm{PIoU}=L_{\mathrm{IoU}}+f(P) \quad 0\leqslant L_{\mathrm{PIoU}}\leqslant 2 \tag{24}$$

$$q=\mathrm{e}^{-P} \quad q\in(0,1],u(x)=3x\cdot\mathrm{e}^{-x^2} \tag{25}$$

$$L_{\mathrm{PIoU}_{at}}=u(\lambda q)\cdot L_{\mathrm{PIoU}}=3\cdot(\lambda q)\cdot\mathrm{e}^{-(\lambda q)^2}\cdot L_{\mathrm{PIoU}} \tag{26}$$

式中：$\mathrm{d}w_1$、$\mathrm{d}w_2$、$\mathrm{d}h_1$和$\mathrm{d}h_2$表示预测框与目标框的对应边之间距离的绝对值；w_{gt}和h_{gt}分别表示目标框的宽度和高度；当预测框和目标框之间完全对齐时，$q=1$，$P=0$，当出现低质量的预测框时，P和q都会增大；λ为注意层超参数；$u(x)$为注意层函数。由于分母是目标框的长宽，所以损失函数的指标因子P不会引起锚箱的扩大，只取决于目标箱的大小。CIoU和PIoU在训练中的目标框和预测框的变化如图5所示。

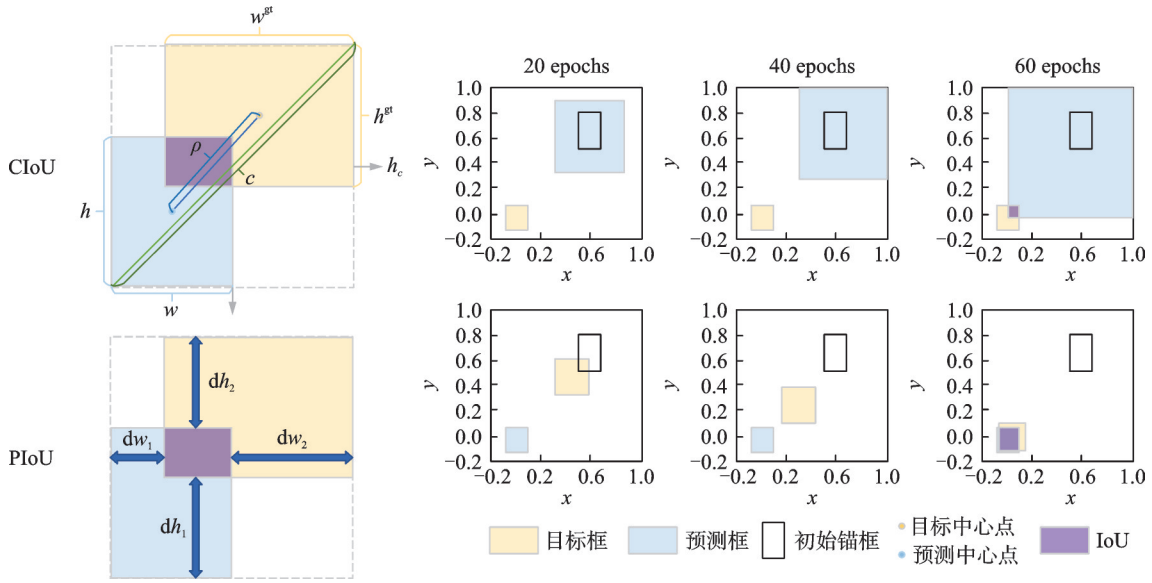

图5 CIoU和PIoU的锚框变化图

Fig.5 Change diagram of anchor frame of CIoU and PIoU

3 实验结果与分析

3.1 数据集和实验环境

本文使用COCO2017人体关键点数据集[21]。COCO2017拥有超过2万张图像数据，其中包含250 000的人体实例，并且每个实例都含有17个关键点注释，其训练集、验证集和测试集分别包括57 000、5 000和20 000张图像。

实验平台为Windows11操作系统，CPU型号为Intel W-2255，GPU型号为GeForce RTX 3090，Py-

thon版本为3.9,CUDA版本为12.2,Pytorch版本为2.2.0。在训练前使用Mosaic方式进行数据增强,模型训练300轮,模型权重大小30 MB。模型训练的超参数设置中,初始学习率为0.01,随机梯度下降的动量为0.937,权重衰减为0.000 5,一共训练300轮,输入图像大小为640像素×640像素,采用平均准确度、参数量及运算量对模型性能进行评价。

3.2 评价指标

为了全面评估模型的识别性能,采用被广泛认可的检测算法评估指标mAP(平均精度均值)、GFlops(GigaFLOPS)和Params(模型参数量)。AP(平均精度)综合考虑准确率和召回率的加权平均值,用于量化模型在某一类别上的总体表现,mAP则是所有类别AP的平均值,旨在评估模型在所有类别上的整体性能。AP_{50}和AP_{75}分别表示当IoU阈值为0.50和0.75时的AP值,AP的计算公式为

$$AP=\int_0^1 P(R)\mathrm{d}R \quad P=\frac{TP}{TP+FP},R=\frac{TP}{TP+FN} \tag{27}$$

式中:TP表示检测器输出中正确检测的样本数,FP表示检测器输出中错误检测的样本数,FN表示真实样本中未被正确检测出的样本数,AP表示精度-召回率曲线下的面积,其大小与网络性能正相关。

3.3 实验与结果分析

将本文提出的PCE_LitePose模型在多人姿态估计方面与主流算法进行对比,这些模型包括DeepPose、OpenPose、EfficientHRNet、YOLOV5s6-Pose以及本文的基础模型YOLOv8s-Pose算法,对比结果如表1所示。

表1 各个算法指标对比

Table 1 Comparison of algorithmic indicators

方法	主干	图像大小/(像素×像素)	参数量/10^6	计算量/10^9	mAP(0.50)/%
DeepPose	ResNet	256×192	23.63		82.4
EfficientHRNet-H1	HRNet	480×480	16.0	28.4	82.6
OpenPose	VGG19	224×224			84.9
YOLOv5s6-Pose	DarkNet-53	640×640	12.6	20.5	82.2
YOLOv8s-Pose	DarkNet-53	640×640	11.6	30.4	85.5
PCE_LitePose (Ours)	**DarkNet-53**	**640×640**	**8.47**	**25.0**	**85.7**

根据表1可知,相比于DeepPose,本文所提模型的参数量降低了64%,准确度上升了4%;相比于EfficientHRNet-H1和OpenPose,其准确度更具有优势。

3.4 消融实验

为了验证本文算法的有效性,本节在YOLOv8Pose的基础上依次添加不同模块进行消融实验验证,结果如表2所示。从表2可以看出,在Backbone部分中添加PCE后,模型参数量和计算量分别下降了12.6%和6%,mAP(0.50)下降0.3%,mAP(0.50:0.95)下降了0.6%;在将Neck部分中添加BiFPN和CARAFE后,参数量和计算量分别上升了2%和1.6%,mAP(0.50)上升了0.5%,mAP(0.50:0.95)上升了1%;在将Head部分引入PSWH后,模型参数量下降了17%,计算量下降了22%,mAP(0.50)不变,mAP(0.50:0.95)下降了1.1%;最后,将损失函数替换为$PIoU_{at}$后,模型的参数量和复杂度保持不变,mAP(0.50:0.95)提升了0.2%,mAP(0.50)上升了0.5%。为了直观比较算法的预测效果,将原始模型添加不同模块后的性能变化过程可视化,包括每一步的参数量(Params)、计算量(GFLOPs)和精度($AP_{0.50:0.95}$和$AP_{0.50}$),结果如图6所示。

表 2 消融实验结果

Table 2 Ablation experimental results

方法	mAP(0.50:0.95)/%	mAP(0.50)/%	参数量/10^6	计算量/10^9
Baseline	58.2	85.5	11.62	30.4
PCE	57.3	85.2	10.15	28.5
BiFPN	58.5	85.8	11.86	30.9
PSWH	57.1	85.4	9.56	26.6
PCE+PSWH	56.2	85.1	8.30	24.6
PCE+BiFPN+PSWH	56.4	85.3	8.47	25.0
PCE+BiFPN+PSWH+$PIoU_{at}$	56.7	85.7	8.47	25.0

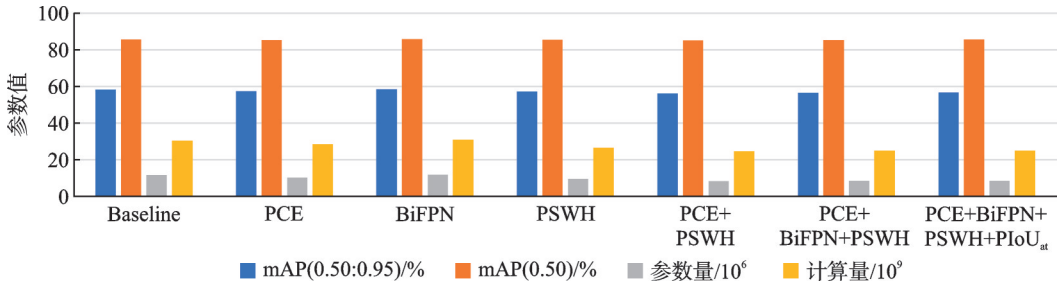

图 6 消融实验可视化结果

Fig.6 Visualization results of ablation experiments

为展示姿态估计的识别效果，选择正常环境下与符合多人、人体遮挡以及小目标条件的 3 组照片进行对比，结果如图 7 所示。

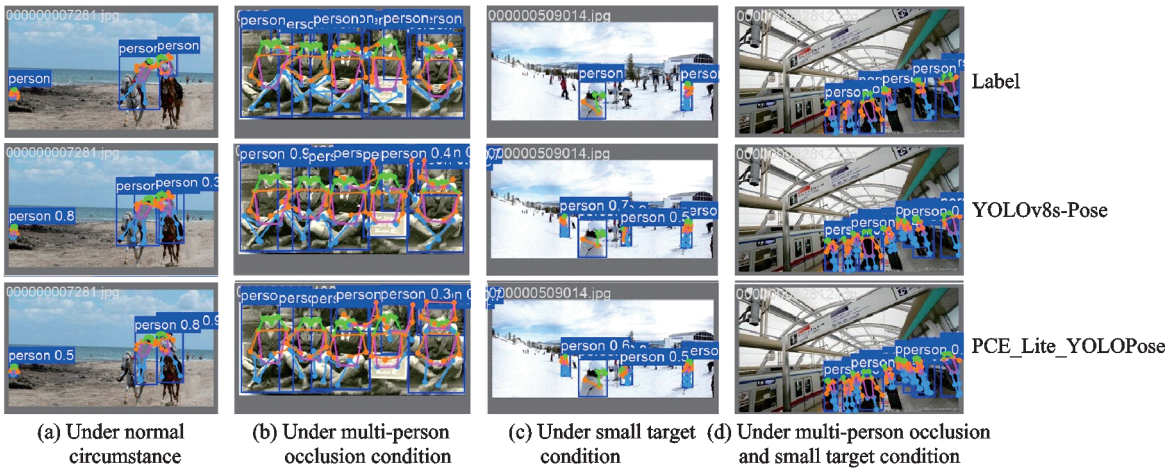

图 7 算法识别结果对比

Fig.7 Comparison of algorithm recognition results

在图 7(a)中，PCE_Lite_YoloPose 的识别结果与标签值相符合，模型也能对缺失部位的关键点进行连接，但是 YOLOv8s-Pose 模型却识别出 4 个人物；在图 7(b)中，标签中并没有标记出关键点的区域，PCE_Lite_YOLOPose 和 YOLOv8s-Pose 都识别出来并进行了连接，但是本文模型识别出了右上角人

物的肩部关节点;图7(c)中面对多人单一背景图片,也能达到很好的识别效果;图7(d)中改进后的模型能准确识别出车站前多人遮挡的人体关键点,经过可视化分析,PCE_Lite_YOLOPose的参数量和计算量更少,对多人姿态估计识别效果更出色。

4 结束语

针对当前姿态估计模型存在参数量大、计算复杂度高的问题,本文提出了一种基于YOLOv8Pose的轻量级网络模型PCE_Lite_YOLOPose,在主干网络中设计了一种混合编码器结构PCE,分别使用卷积神经网络和视觉编码器处理特征,该模块利用两者的优势实现局部和全局特征提取,在颈部网络使用BiFPN,提升模型的颈部网络的多尺度特征融合能力,将检测头部分替换为局部权重共享头PSWH,通过共享卷积进行人体检测和分类任务,降低了存储需求和计算成本。该模型在参数量、检测速度和准确性等方面都有所提升,在COCO2017数据集上的实验结果显示了该算法在人体姿态估计任务中的优越性。

参考文献:

[1] PISHCHULIN L, INSAFUTDINOV E, TANG S, et al. DeepCut: Joint subset partition and labeling for multi person pose estimation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 4929-4937.

[2] 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. Piscataway, NJ: IEEE, 2017: 7291-7299.

[3] KREISS S, BERTONI L, ALAHI A. PifPaf: Composite fields for human pose estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 11977-11986.

[4] HE K, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE, 2017: 2961-2969.

[5] FANG H S, LI J, TANG H, et al. AlphaPose: Whole-body regional multi-person pose estimation and tracking in real-time[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(6): 7157-7173.

[6] MAJI D, NAGORI S, MATHEW M, et al. YOLO-Pose: Enhancing YOLO for multi person pose estimation using object keypoint similarity loss[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 2637-2646.

[7] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 7132-7141.

[8] WOO S, PARK J, LEE J Y, et al. CBAM: Convolutional block attention module[C]//Proceedings of the European Conference on Computer Vision (ECCV). Berlin: Springer, 2018: 3-19.

[9] 张鹏, 彭宗举, 张文瑞, 等. 多级注意力特征优化的道路场景实时语义分割[J]. 数据采集与处理, 2024, 39(6): 1505-1516.
ZHANG Peng, PENG Zongju, ZHANG Wenrui, et al. Real-time semantic segmentation of road scene based on multi-level attention feature optimization[J]. Journal of Data Acquisition and Processing, 2024, 39(6): 1505-1516.

[10] 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). https://doi.org/10.48550/arXiv.2010.11929.

[11] HOWARD A G, ZHU M L, CHEN B, et al. MobileNets: Efficient convolutional neural networks for mobile vision applications[EB/OL]. (2017-04-17). https://doi.org/10.48550/arXiv.1704.04861.

[12] OSOKIN D. Real-time 2D multi-person pose estimation on CPU: Lightweight OpenPose[EB/OL]. (2018-11-29). https://doi.org/10.48550/arXiv.1811.2004.

[13] TAN M, LE Q. EfficientNet: Rethinking model scaling for convolutional neural networks[C]//Proceedings of International Conference on Machine Learning. New York, USA: PMLR: 6105-6114.

[14] LI Q, ZHANG Z, XIAO F, et al. Dite-HRNet: Dynamic lightweight high-resolution network for human pose estimation[EB/OL]. (2022-05-22). https://doi.org/10.48550/arXiv.2204.10762.

[15] 贾子豪, 王文青, 刘光灿. 改进YOLOv5的轻量化交通标志检测算法[J]. 数据采集与处理, 2023, 38(6): 1434-1444.
JIA Zihao, WANG Wenqing, LIU Guangcan. Improved lightweight traffic sign detection algorithm of YOLOv5[J]. Journal of Data Acquisition and Processing, 2023, 38(6): 1434-1444.

[16] TAN M, PANG R, LE Q V. Efficientdet: Scalable and efficient object detection[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 10781-10790.

[17] WANG J, CHEN K, XU R, et al. Carafe: Content-aware reassembly of features[C]//Proceedings of IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2019: 3007-3016.

[18] LI X, WANG W, WU L, et al. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection[J]. Advances in Neural Information Processing Systems, 2020, 33: 21002-21012.

[19] ZHENG Z, WANG P, REN D, et al. Enhancing geometric factors in model learning and inference for object detection and instance segmentation[J]. IEEE Transactions on Cybernetics, 2021, 52(8): 8574-8586.

[20] LIU C, WANG K, LI Q, et al. Powerful-IoU: More straightforward and faster bounding box regression loss with a nonmonotonic focusing mechanism[J]. Neural Networks, 2024, 170: 276-284.

[21] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: Common objects in context[C]//Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland: Springer, 2014: 740-755.

作者简介:

徐新智(2000-),男,硕士研究生,研究方向:姿态估计、动作识别,E-mail:222302312@st.usst.edu.cn。

何宏(1973-),**通信作者**,女,教授,研究方向:医学人工智能、人机交互系统、医疗大数据分析,E-mail:hehong@usst.edu.cn。

(编辑:王静)