基于自适应参数化非极大值抑制的二维人体姿态估计算法

李家宝[1,2]，王成军[1,2]，苏文杭[1]

（1.安徽理工大学 人工智能学院，安徽 淮南 232001；2.合肥综合性国家科学中心 人工智能研究院，合肥 230026）

摘 要：针对检测器的准确率低和检测结果导致的关键点冗余两个问题，提出了基于自适应参数化NMS的二维人体姿态估计算法。首先，采用CenterNet检测器取代原有检测器，提高了人体检测的性能，并为后续姿态估计打下基础。然后，提出了自适应参数化的PoseNMS算法，引入样本外观相似度实现样本自适应度量调节，使NMS中的过滤条件更加灵活。最后，提出了一种基于检测置信度的非均匀采样方法，在训练过程中保证了样本的有效性，实现了困难样本的发现与挖掘。在MSCOCO 2017数据集上采用本文检测器输出的检测结果作为后续目标框的条件取得了71.9%的mAP。另外，本文算法同样在MPII和MSCOCO 2015数据集上进行了大量实验，定量实验与可视化结果说明本文方法有效解决了上述两个问题，实现了更准确的姿态估计。

关键词：CenterNet；关键点检测；非极大值抑制；姿态估计；非均匀采样

中图分类号：TP391 **文献标志码**：A **文章编号**：1671-5497(2025)07-2425-09

DOI：10.13229/j.cnki.jdxbgxb.20240744

2D human pose estimation algorithm based on adaptive parameterized non-maximum suppression

LI Jia-bao[1,2], WANG Cheng-jun[1,2], SU Wen-hang[1]

(1. *School of Artificial Intelligence, Anhui University of Science and Technology, Huainan* 232001, *China*; 2. *Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei* 230026, *China*)

Abstract: To address the two problems of low detector accuracy and redundant keypoints in detection results, a two-dimensional human pose estimation algorithm was proposed based on adaptive parameterized NMS. Replacing the original detector with CenterNet detector improves the performance of human detection and lays the foundation for subsequent pose estimation. Propose an adaptive parameterized PoseNMS algorithm that introduces sample appearance similarity to achieve sample adaptive measurement adjustment, making the filtering conditions in NMS more flexible. A non-uniform sampling method based on detection confidence was proposed, which ensured the effectiveness of the samples during the training

收稿日期：2024-07-05.

基金项目：安徽省高校协同创新项目(GXXT-2022-053)；国家创新方法工作专项项目(2018IM010500)；淮南市科技计划项目(2021A242).

作者简介：李家宝(1998-)，男，博士研究生.研究方向：人体姿态估计，人工智能.E-mail：3196177554@qq.com

process and achieved the discovery and mining of difficult samples. Verified on three datasets, a 71.9% mAP was achieved on the MSCOCO 2017 dataset under the condition of using the detection results output by the detector in this paper as the subsequent target box. In addition, the algorithm proposed in this paper has also been extensively experimented on MPII and MSCOCO 2015, and the quantitative and visual results show that the proposed method effectively solves the above two problems and achieves more accurate attitude estimation.

Key words: CenterNet; key point detection; non-maximum suppression; pose estimation; non-uniform sampling

0 引 言

随着计算机视觉技术的不断发展和深度学习技术的广泛应用,二维人体姿态估计成为计算机视觉领域的一个重要研究方向[1]。该研究旨在从图像或视频中定位并识别出人体的关键点,通过将这些关键点按顺序相连构建出人体的躯干,从而实现对人体姿态的估计。这一技术不仅有助于动作识别、人机交互和姿态跟踪等,还在自动驾驶、安全监控、特殊场景下的动作监控等领域具有潜在的应用前景。然而,二维人体姿态估计仍面临诸多挑战,如拥挤场景下的目标分离、遮挡问题以及实时性要求等,这些问题的解决需要不断地研究和优化算法。

在深度学习等技术快速发展的背景下,二维人体姿态估计算法的研究现状呈现出快速发展的趋势。其中,马皖宜等[2]提出了基于多谱注意力高分辨率网络的人体姿态估计方法,融合多谱注意力与Lite-HRNet,利用频率分量提取丰富的特征信息,通过通道置换、分组卷积与深度可分离卷积,提升姿态估计速度。虽然多谱注意力机制和多分辨率特征融合有助于提取更丰富的特征信息,但在具体实现过程中,如何更有效地融合不同尺度的特征,以及如何处理特征融合过程中的信息丢失问题,仍是该方法需要进一步优化和改进的方向。吉斌等[3]提出一种时空信息感知网络STNet,STNet采用卷积模块提取视频帧人体关节点热力图,循环卷积模块编码时间信息,通过时空解耦学习提升姿态估计连贯性与准确性。由于STNet网络结构相对复杂,参数调优和训练过程可能较为困难,如果参数设置不合理,会造成估计结果不精准。Cao等[4]提出了实时多人2D姿势估计法,将非参数部分亲和场(PAFs)结合身体与足部检测器,实现快速准确的关键点定位。该方法依赖于身体部位位置的初步检测结果来构建部分亲和场(PAFs),如果初始检测结果存在误差,这种误差可能会传递到后续的关键点关联步骤中,从而影响整体的姿态估计结果。

针对现有研究存在的问题,本文提出了一个基于准确检测和自适应参数化NMS的两阶段二维姿态估计算法:①采用先进的CenterNet检测器,相比于原有检测器,CenterNet采用直接回归中心点与偏移量的方式进行检测,提高了目标检测精度,为后续姿态估计奠定基础;②由于传统PoseNMS对相似结构关键点与相近关键点采用预定义的度量方式,无法自适应判断当前关键点的外观与结构。因此,本文提出自适应参数化的Pose NMS方法,通过引入人体外观相似度作为度量条件,使NMS中的过滤条件更加灵活;③在训练中,为了保证样本的有效性,本文提出基于检测置信度的非均匀采样方法。算法基于先前检测器的置信度分数,计算样本采样概率,实现困难样本的挖掘。

本文算法在MSCOCO 2017测试集上进行实验对比,得到了71.9%的mAP(平均准确率)。比基准算法提升了0.9%。另外,本文算法还在MPII和MSCOCO 2015测试集上进行了可视化实验。其结果表明:本文算法能够有效应对小目标、光照受限、形变、目标遮挡、半身与多人等场景。

1 二维人体姿态估计算法设计

本文算法框架如图1所示。首先,引入高效的人体检测器CenterNet,提升目标检测精度,为后续人体姿态估计提供更准确的目标框。在测试过程中,将得到的人体候选框输入姿态估计器(Pose estimator)中,得到初始结果。为了得到更准确的人体姿态,提出了自适应参数化的Po-

seNMS(Adaptive PoseNMS)模块，同时考虑关键点距离与位姿相似度，并引入外观相似度解决两种度量不一致的问题，得到最终的二维人体姿态。在训练过程中，为了更好地根据检测框生成候选目标框，本文提出基于检测置信度的非均匀采样方法(Confidence-guided non-uniform sampling)，为后续位姿估计器的训练生成候选框，挖掘复杂样本，实现有效训练。

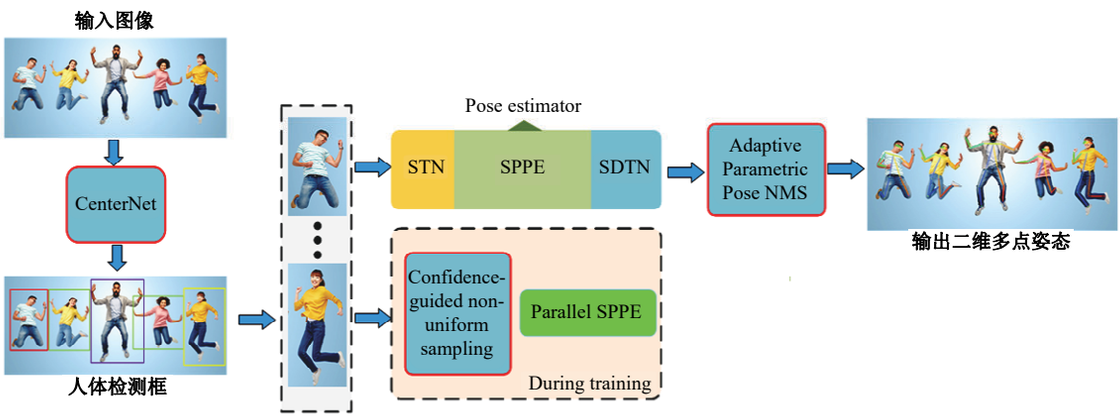


图 1　网络结构

Fig. 1　Network structure

1.1　AlphaPose

AlphaPose 是一种区域多人姿态估计框架，采用的是自顶向下的方法，但目前这种方法存在下面两个问题，即边界框定位不准确和姿态冗余问题。基于上述问题，引入一种区域多人姿势估计(Regional multi-person pose estimation，RMPE)框架，该框架由 3 个模块组成，分别为人体检测器/姿态估计器和参数化姿态 NMS[5-7]。

AlphaPose 网络框架如图 2 所示。首先 AlphaPose 采用 YOLO 检测器提供目标检测结果，以进一步提高精度。AlphaPose 将得到的人体检测框送入姿态估计器中，其中姿态估计器包括 STN(Spatial transformer network)、SPPE(Single person pose estimation)和 SDTN(Spatial de-transformer network)。SPPE 表示单人姿态估计模块，其输入为单人人体检测框，输出一个人体姿态。为了消除 SPPE 对人体位置造成的检测影响，算法在 SPPE 前增加了 STN 模块对检测器得到的检测框进行空间变换，消除定位误差的影响[8,9]。之后，使用 SDTN 模块对人体位置进行反变换，将得到的姿态结果映射回原始图片坐标。为了帮助 STN 更好地提取人体区域，算法还并联了一个 SPPE 支路，并且共享 STN 的结果。针对姿态冗余问题，提出了参数化姿态非最大抑制(Parametric PoseNMS)，其中包含两个抑制度量函数即置信度度量和距离度量，其方法如下：

将具有 m 个关节的位姿 p_i 表示为 $\left\{\left\langle k_i^1, c_i^1\right\rangle, \cdots, \left\langle k_i^m, c_i^m\right\rangle\right\}$，其中，$k_i^j$ 和 c_i^j 分别为关节的第 j 个位置和置信度得分。其方案为选择最可信的位姿作为参考，一些接近它的位姿可以通过 NMS 去除冗余位姿。在剩余的姿势集上重复此过程，直到消除多余的位姿。因此，需要定义位姿相似性，以消除彼此接近和相似的位姿[10,11]。定义一个位姿距离度量 $d\left(P_i, P_j|\Lambda\right)$ 来测量位姿相似性，并定义一个阈值 η 作为判断标准，其中，Λ 为函数 $d(\cdot)$ 的参数集。判断标准如下：

$$f\left(P_i, P_j|\Lambda, \eta\right)=1\left[d\left(P_i, P_j|\Lambda, \lambda\right)\leqslant\eta\right] \quad (1)$$

判断标准包括置信度度量和距离度量，其中，置信度度量是指消除置信度相似的关节点，选择置信度最高的姿态 P_i 作为参考姿态，判断姿态是否需要被消除，置信度度量如式(2)所示：

$$K_{\text{sim}}\left(P_i, P_j|\sigma_1\right)= \begin{Bmatrix} \sum_n \tanh\dfrac{c_i^n}{\sigma_1}\cdot\tanh\dfrac{c_j^n}{\sigma_1}, \text{if } k_j^n \text{ within } B\left(k_i^n\right) \\ 0, \quad \text{otherwise} \end{Bmatrix} \quad (2)$$

式(3)用来计算两个姿态关节点的空间距离总和，选取置信度较高的姿态 P_i，如果 P_j 姿态与 P_i 距离比较近，这说明两者重叠度较高，那么 P_j 就会被消除：

$$H_{\text{sim}}\left(P_i, P_j|\sigma_2\right)=\sum_n \exp\left[-\frac{\left(k_i^n-k_j^n\right)^2}{\sigma_2}\right] \quad (3)$$

由此可得到距离函数为:

$$d(P_i,P_j|\Lambda)=K_{\text{sim}}(P_i,P_j|\sigma_1)+\lambda H_{\text{sim}}(P_i,P_j|\sigma_2) \quad (4)$$

式中:λ为平衡两个距离的权重,并且$\Lambda=\{\sigma_1,\sigma_2,\lambda\}$。

在训练过程中,算法通过数据增强来训练模块。一种方法是利用检测到的人体框进行训练,但目标检测通常只产生单个人体框,限制了STN和SPPE模块的充分训练。因此,引入姿态引导区域框生成器(Pose-guided proposals generator,PGPG)进行数据增强,通过模拟不同姿态的人体检测器输出分布,来生成大量训练样本。

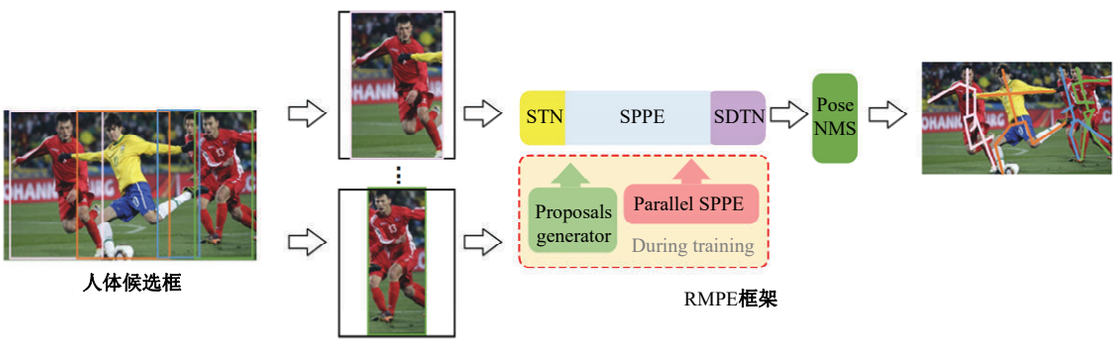


图2 区域多人姿态估计(RMPE)框架

Fig. 2 Regional multi person attitude estimation(RMPE) framework

由于两阶段位姿检测器首先需要对全图进行目标检测,因此,目标检测结果对后续操作起关键性作用。AlphaPose采用单阶段基于锚框的检测器YOLO系列提供目标检测结果,而基于锚框的检测器相比于不使用锚框的检测器而言,每个锚框都需计算两个框之间的交并比,导致运算复杂度增大,产生的参数较多。因此,首先对AlphaPose方法中的目标检测器进行改进,使用先进的CenterNet检测器替代原有检测器。

CenterNet构建模型时将目标作为一个点,将图像输入网络后,会生成关键点热图,热图上的高峰点即目标中心。检测到中心点后,根据每个高峰点的特征回归出目标框的其他属性,包括二维框的宽高、三维框的长宽高、深度、角度及人体姿态等。CenterNet框架如图3所示。最左边表示输入图片,输入图片需要裁剪到512×512大小,即长边缩放到512,短边补0。中间表示基准网络,采用了Hourglass、ResNet与DLA 3种网络架构。其优势在于:①锚点仅仅是放在目标中心,没有尺寸框;②使用更大分辨率的输出特征图,有助于改善小目标检测质量。

具体来说,首先对图像进行预处理以满足目标检测算法需求。第一步需要将图像进行归一化,接下来,通过CenterNet输出检测到的目标中心点热力图、当前检测目标的置信度以及目标框尺度回归值,并确定各目标的检测框。由于CenterNet只估计目标的中心点位置,因此,检测算法中不涉及NMS等后处理操作,最后根据目标框位置,对检测到目标区域进行裁剪,用于后续关键点估计算法。

1.2 自适应参数化位姿关键点非极大值抑制

基于AlphaPose中的参数化NMS方法存在度量不平衡的问题,即可能存在冗余区域与其中一个度量方法类似,而与另一种度量方法区别较大,导致抑制失败。发现上述两种度量方法是互补关系,不一定同时起作用。因此,本文提出一种自适应参数化的PoseNMS方法,使用特征提取网络提取目标框特征并计算逐像素特征相似度,用于抑制相似结构的检测框,使NMS中的过滤条件更加灵活。

其中网络以区域图像、区域关键点、候选位姿作为输入,经过VGG16网络提取特征,并使用全局特征池化进行特征压缩,最后使用全连接层及Sigmoid函数输出权重。将P_i和P_j所在图像块$I(P_i)$、$I(P_j)$作为输入之后,并将输出的外观特征相似度$A_{\text{sim}}\left(\left(I(P_i),I(P_j)\right)|\varphi\right)$,作为度量条件,得到一个新的自适应参数化距离函数:

$$d_{\text{adaptive}}(P_i,P_j|\Lambda)=\frac{K_{\text{sim}}(P_i,P_j|\sigma_1)+H_{\text{sim}}(P_i,P_j|\sigma_2)}{A_{\text{sim}}\left(I(P_i),I(P_j)|\varphi\right)} \quad (5)$$

具体实现如下:首先将检测器给出的目标区

域作为输入，定义 VGG 16特征提取网络φ提取其特征，并利用全局平均池化将特征聚合成$1\times 1\times c$的向量，并将两个聚合向量进行逐元素相乘然后，将结果相加，得到两个框的相似度$\left(I_{\text{sim}}\left(P_i,P_j|\varphi\right)\in(0,1]\right)$并作为现有距离函数的系数，即相似度越高的框，其距离越短，越可能为冗余框；反之亦然。其网络框架如图 4 所示。其意义在于通过引入外观相似度自适应地调整 NMS 的距离，更好地去除冗余关键点和候选框，提升关键点检测精度。

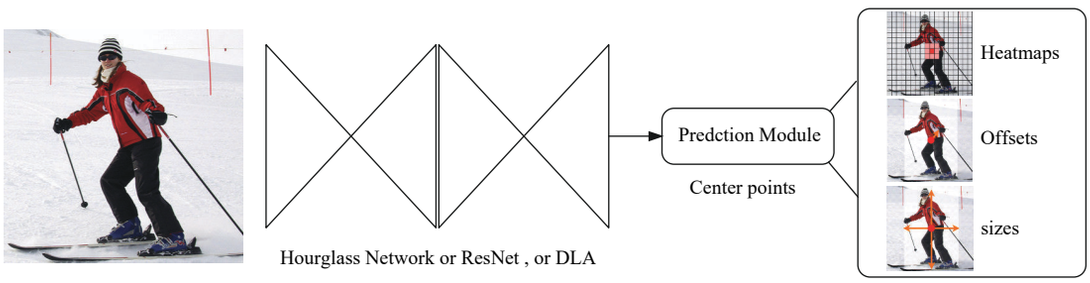


图 3　CenterNet 网络的框架图

Fig. 3　Framework diagram of CenterNet network

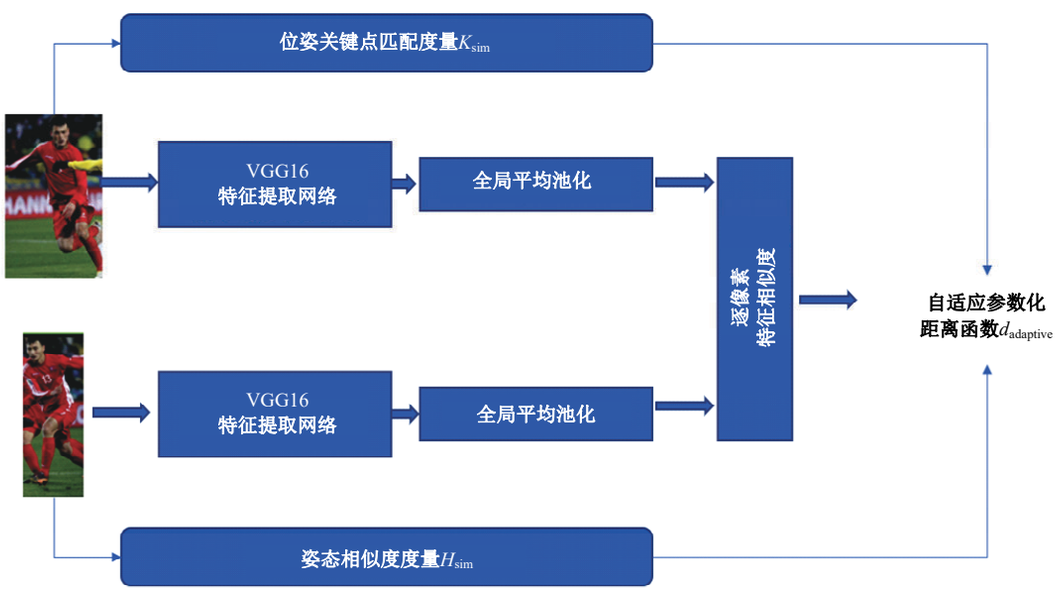


图 4　自适应位姿距离度量网络框架图

Fig. 4　Framework diagram of adaptive pose distance measurement network

1.3　基于检测置信度的非均匀采样方法

对于两阶段姿态估计，适当的数据增强有助于让 STN＋SPPE＋SDTN 适应人体区域定位结果，AlphaPose 采用姿态引导区域框生成器（PGPG）进行数据增强，PGPG 是类似高斯分布的采样方法，即在检测框周围根据距离目标框的位置进行采样，上述方法根据距离进行判断，在目标框检测出现偏差时，目标区域相对于背景区域可能出现相同权重，这种方法由于不包含图像信息，因此，可能会出现不包含人体位置的区域，显然这种情况在检测时是不合理的。

因此，提出了利用之前检测器的置信度，用检测器的结果预先筛选目标，其意义在于：可以利用检测器的检测置信度，对目标可能出现的区域进行初步筛选，提高候选区域采样的效率和准确性。与现有方法对比如图 5 所示。

具体流程如下：首先生成以目标为中心、以目标宽高为参数的二维高斯分布，并且考虑检测器的 heatmap 结果，从而生成采样概率P_{ij}，其中采样概率如下：

$$P_{ij}=p_{ij}+\text{Norm}\left(H_{ij}\right) \tag{6}$$

$$\text{Norm}(H)=\frac{H-\min(H)}{\max(H)-\min(H)} \tag{7}$$

式中：p_{ij}服从二维高斯分布；H为检测器的 Heatmap 图。

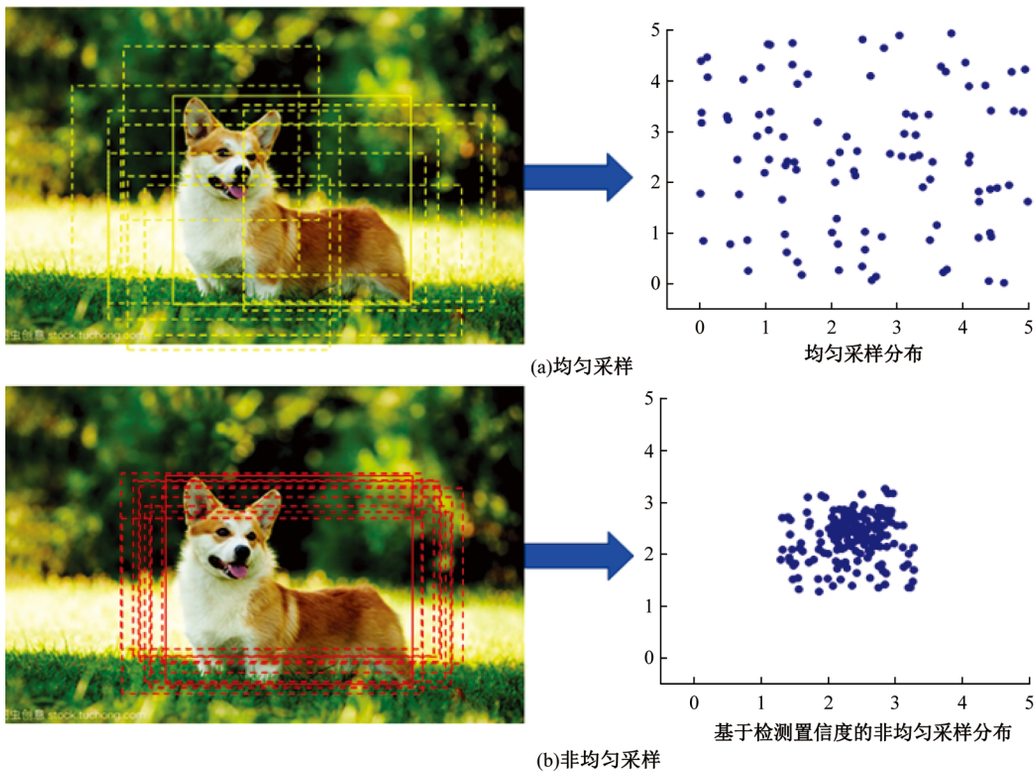


图5 基于检测置信度的非均匀采样方法与高斯分布的采样方法对比

Fig. 5 Comparison between non-uniform sampling method based on detection confidence and sampling method based on Gaussian distribution

2 实验验证

为了验证本文方法的有效性,展开实验研究。实验中使用的硬件配置:CPU为Intel® Core™ 19-13900KS 3.20 GHz,GPU为NVIDIA GeForce RTX 4080(24 GB),编译器为Python 3.9。实验使用Anaconda 3深度学习模型开发工具和PyTorch 1.10.1构建网络模型。

2.1 评估数据集

2.1.1 MPII

MPII Human Pose数据集是一个广泛使用的大规模多视图2D人体姿态估计数据集,该数据集涵盖4万个标注了16个关键点信息的人体目标,这些关键点的准确标注对人体姿态估计至关重要,如图6所示。在本文中,算法使用单帧多人姿态测试集作为实验数据集,选取14 679张图片用于训练,6 619张用于测试,2 729张用于验证。

2.1.2 COCO 2015

MSCOCO 2015数据集是一种用于计算机视觉研究的数据集,其中包含了100多万张图像和32万个与之相对应的注释,来源于250 000张电影截图和328 000张现实图像。COCO 2015中的关键点包括15个部位,分别是:鼻子、两眼、两耳、两肩、两肘、两手腕、两髋、两膝、两脚踝。由于其包含了丰富多样的图像和详细精准的注释信息,因此,

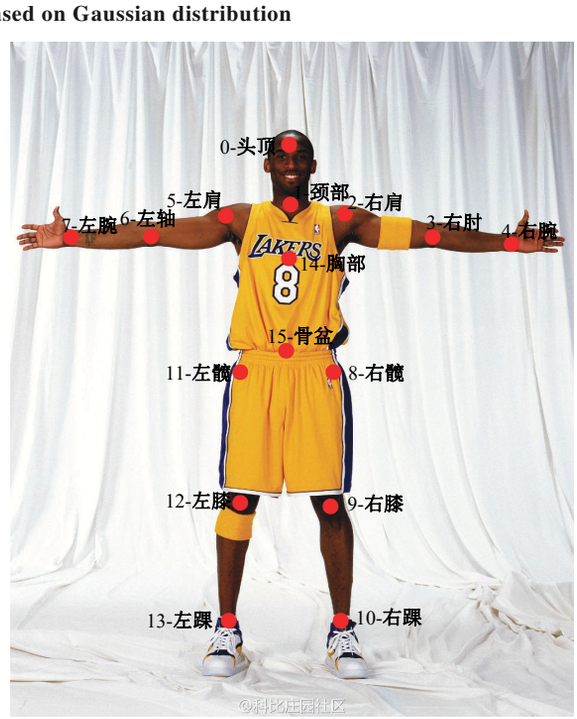


图6 MPII数据集关键点描述

Fig. 6 Description of key points in MPII dataset

对训练和评估深度学习模型具有很高的实用性。

2.1.3　COCO 2017

MSCOCO 2017 数据集是 COCO（Common objects in context）标注协议下的一个升级版本，它是目前计算机视觉领域中使用最广泛的一种数据集。该数据集包括超过 330 k 张图像、2.5 M 个实例，其中包括在 22 k 个不同场景下拍摄的日常图像和专业拍摄的艺术图片。MSCOCO 2017 在 MSCOCO 2015 数据集上进一步扩展，增加了 5 个关键点，包括脊椎、嘴唇、左眉毛、右眉毛、舌头。这样 MSCOCO 2017 中的关键点一共有 20 个，相比 MSCOCO 2015 更加细致。

2.2　评估指标

平均精度均值 mAP 通常用来评估目标检测算法的性能，它是不同阈值下精度与召回率的面积积分的平均值。AP 值是对每一个类别通过 Precision 和 Recall 的取值来计算出来的，计算公式为：

$$\mathrm{mAP}=\frac{1}{|classes|}\sum_{c\in classes}\frac{|\mathrm{TP}_c|}{|\mathrm{FP}_c|+|\mathrm{TP}_c|} \tag{8}$$

式中：TP_c 为准确检测；FP_c 为虚检。

2.3　mAP 结果

将文献[2]方法、文献[3]方法和文献[4]方法作为比较算法，对比不同方法的 mAP，结果如表 1 所示。

表 1　COCO2017 测试集上的对比结果说明

Table 1　Comparison results on COCO 2017 test set

方法	mAP	模型参数量/M	推理时间/ms
文献[2]方法	61.8%	25	150
文献[3]方法	65.5%	25	75
文献[4]方法	71.0%	30	25
本文方法	71.9%	30	22

如表 1 所示，本文方法利用 CenterNet 作为检测器，取得了 71.9% 的 mAP 结果，对比其他方法具有显著的改进。与第二名的文献[4]方法相比，本文方法在 mAP 上提高了 0.9%。同时，参数数量和推理时间是模型性能的两个关键指标。参数数量表示模型的复杂性，而推理时间反映了模型的操作速度。从表 1 可以看出，虽然本文方法的参数数量（30 M）超过了文献[2]方法和文献[3]方法。但文献[2]方法和文献[3]方法的推理时间分别为 150 ms 和 7 5 ms，均多于本文方法。这是因为本文方法结合了 CenterNet 检测器、APPoseNMS，实现了令人满意的推理速度，推理时间为 22 ms，在推理时间方面排名第一。由此可知，本文方法在 2D 人体姿态估计算法领域展现出了出色的检测准确性和推理速度，使其适用于各种实际场景。

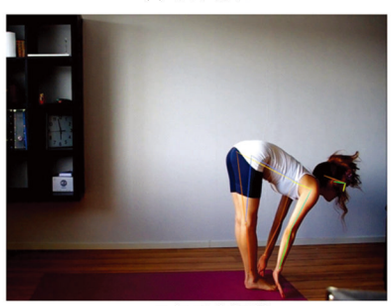
2.4　可视化结果

2.4.1　本文方法在 MPII 测试集上的可视化结果

本文方法在 MPII 测试集上的可视化测试结果如图 7 所示。

(a) 多人场景

(b) 半身场景



(c) 目标形变场景

图 7　本文方法在 MPII 数据集上的可视化结果

Fig. 7　Visualization results of proposed method on MPII dataset

本文方法能够有效处理以下问题：在多人场景中，出现目标被部分遮挡的情况时，本文方法可以较好地处理这些场景；在半身场景中，目标部分关键点缺失场景会导致姿态估计算法无法匹配全部关键点，因而造成关键点误匹配的问题，本文方法由于适应了高效检测器可以解决这些问题；对于目标形变场景，由于目标发生了非刚性形变，导致目标形状改变，加大了检测难度。另外，算法可能学习到关键点具有相对位置关系，导致检测形变物体时出现误差。本文方法提出了自适应参数化的 Po-

seNMS,可以对冗余关键点进行自适应距离度量,因而可以在上述挑战场景中得到较好的结果。

2.4.2 在MSCOCO 2015测试集上

本文方法在MSCOCO 2015测试集上的可视化测试结果如图8所示。

对于目标遮挡场景,由于目标之间存在交叉遮挡的现象,很容易降低关键点匹配的准确性,本文方法能够有效解决此问题。

(a) 目标遮挡场景 (b) 半身场景 (c) 多人场景

图8 本文方法在MSCOCO 2015测试集上的可视化结果

Fig. 8 Visualization results of proposed method on MSCOCO 2015 test set

2.4.3 在MSCOCO 2017 测试集上

本文方法在MSCOCO 2017测试集上的可视化测试结果如图9所示。在小目标场景下,目标的特征点往往比较微小,很容易受到噪声的干扰,在训练中,采用基于检测置信度的非均匀采样方法进行数据增强,增强模型的泛化能力和适应性。在光照受限的场景中,由于光线的变化,目标的视觉特征往往会失真或丢失,导致算法无法正确地估计目标姿态,采用本文方法可有效解决。

(a) 小目标场景 (b) 光照受限场景 (c) 目标形变场景

图9 本文方法在MSCOCO 2017测试集上的可视化结果

Fig. 9 Visualization results of proposed method on MSCOCO 2017 test set

3 结束语

首先,本文提出了一种基于自适应参数化NMS的二维人体姿态估计算法,采用CenterNet检测器来替代原始检测器,不仅提高了人体检测的性能,还为后续的姿态估计奠定了坚实的基础。其次,本文通过提出自适应参数化的PoseNMS方法,增强了原始的PoseNMS方法,可以根据候选样本外观的相似性自适应生成判断权重,从而更加灵活地调整NMS中的过滤条件。最后,为了确保训练样本的有效性,本文还提出了一种基于检测置信度的非均匀采样方法,可以在训练过程中发现和挖掘困难样本。本文在MSCOCO 2017数据集上进行的定量验证取得了71.9%的mAP。此外,本研究在MPII和MSCOCO 2015测试集上对该方法进行的视觉描述显示,在诸如光照有限和目标较小等各种具有挑战性的情况下,该方法能够取得显著的效果。在未来的工作中,计划利用本文所提出的二维人体姿态估计方法进一步开发更先进的三维人体姿态估计技术。此外,还计划创建更具针对性的数据集,其中包含更大量的数据,专门在步态分析和医疗康复领域等特定场景中应用,人体姿态估计方法将为医生提供宝贵的参考资源。

参考文献:

[1] 张宇, 温光照, 米思娅, 等. 基于深度学习的二维人体姿态估计综述[J]. 软件学报, 2022, 33(11): 4173-4191.

Zhang Yu, Wen Guang-zhao, Mi Si-ya, et al. A re-

view of two-dimensional human pose estimation based on deep learning[J]. Journal of Software, 2022, 33(11): 4173-4191.

[2] 马皖宜, 张德平. 基于多谱注意力高分辨率网络的人体姿态估计[J]. 计算机辅助设计与图形学学报, 2022, 34(8): 1283-1292.
Ma Wan-yi, Zhang De-ping. Human pose estimation based on multi-spectral attention high-resolution network[J]. Journal of Computer-Aided Design and Computer Graphics, 2022,34(8):1283-1292.

[3] 吉斌, 潘烨, 金小刚, 等. 用于视频流人体姿态估计的时空信息感知网络[J]. 计算机辅助设计与图形学学报, 2022, 34(2): 189-197.
Ji Bin, Pan Ye, Jin Xiao-gang, et al. Spatio-temporal information perception network for human pose estimation in video streams[J]. Journal of Computer-Aided Design and Computer Graphics, 2022, 34(2): 189-197.

[4] Cao Z, Hidalgo G, Simon T, et al. Openpose: realtime multi-person 2D pose estimation using part affinity fields[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021 43(1):172-186.

[5] 程旭, 宋晨, 史金钢, 等. 基于深度学习的通用目标检测研究综述[J]. 电子学报, 2021, 49(7): 1428-1438.
Cheng Xu, Song Chen, Shi Jin-gang, et al. A review of general object detection based on deep learning[J]. Acta Electronica Sinica, 2021, 49(7): 1428-1438.

[6] 邵延华, 张铎, 楚红雨, 等. 基于深度学习的YOLO目标检测综述[J]. 电子与信息学报, 2022, 44(10): 3697-3708.
Shao Yan-hua, Zhang Duo, Chu Hong-yu, et al. A review of YOLO object detection based on deep learning[J]. Journal of Electronics & Information Technology, 2022, 44(10): 3697-3708.

[7] 陈丽, 王世勇, 高思莉, 等. Sentinel-2卫星的多光谱轻量级船舶目标检测算法[J]. 光谱学与光谱分析, 2022, 42(9): 2862-2869.
Chen Li, Wang Shi-yong, Gao Si-li, et al. Multispectral lightweight ship detection algorithm for Sentinel-2 satellite[J]. Spectroscopy and Spectral Analysis, 2022, 42(9): 2862-2869.

[8] 陈科圻, 朱志亮, 邓小明, 等. 多尺度目标检测的深度学习研究综述[J]. 软件学报, 2021, 32(4): 1201-1227.
Chen Ke-qi, Zhu Zhi-liang, Deng Xiao-ming, et al. A review of deep learning research on multi-scale object detection[J]. Journal of Software, 2021, 32(4): 1201-1227.

[9] Law H, Deng J. CornerNet: detecting objects as paired keypoints[J]. International Journal of Computer Vision, 2020, 128(3): 642-656.

[10] 曲优, 李文辉. 基于锚框变换的单阶段旋转目标检测方法[J]. 吉林大学学报:工学版, 2022, 52(1): 162-173.
Qu You, Li Wen-hui. Single-stage rotated object detection method based on anchor box transformation [J]. Journal of Jilin University(Engineering and Technology Edition), 2022,52(1): 162-173.

[11] 朱毅琳, 肖秦琨, 杨梦薇. 基于语义图卷积神经网络的三维人体姿态估计[J]. 计算机仿真, 2023, 40(11): 207-211, 402.
Zhu Yi-lin, Xiao Qin-kun, Yang Meng-wei. 3D human pose estimation based on semantic graph convolutional neural network[J]. Computer Simulation, 2023, 40(11): 207-211, 402.