基于视频的实时多人姿态估计方法

闫芬婷，王鹏*，吕志刚，丁哲，乔梦雨
西安工业大学电子信息工程学院，陕西 西安 710021

摘要 针对图像和视频中多人姿态估计存在人体边界框定位不准确、困难关键点检测精度有待提高等问题，设计了一套基于自顶向下框架的实时多人姿态估计模型。首先将深度可分离卷积加入目标检测算法中，提高人体检测器运行速度；然后基于特征金字塔网络结合上下文语义信息，采用在线难例挖掘算法解决困难关键点检测精度低的问题；最后结合空间变换网络与姿态相似度计算，剔除冗余姿态，改善边界框定位准确性。本文提出模型在 2017MS COCO Test-dev 数据集上的平均检测精度比 Mask R-CNN 模型提升了 14.84%，比 RMPE 模型提升了 2.43%，帧频达到 22 frame/s。

关键词 图像处理；多人姿态估计；空间变换网络；语义信息；姿态距离
中图分类号 TP391.4 **文献标志码** A **doi**：10.3788/LOP57.021006

Real-Time Multi-Person Video-Based Pose Estimation

Yan Fenting, Wang Peng*, Lü Zhigang, Ding Zhe, Qiao Mengyu
School of Electronics and Information Engineering, Xi'an Technological University, Xi'an, Shaanxi 710021, *China*

Abstract For multi-person pose estimation in images and videos, it is necessary to address the inaccurate positioning of the human-bounding box and improve the detection accuracy of hard keypoints. This paper designs a real-time multi-person pose-estimation model based on a top-down framework. First, depth-separable convolution is added to the target-detection algorithm to improve the running speed of the human detector; then, by combining the feature pyramid network with context-semantic information, the online hard-example mining algorithm is used to solve the problem of low detection accuracy at hard keypoints. Finally, combining the spatial-transformation network and pose-similarity calculation, the redundant pose is eliminated and the accuracy of the bounding-box positioning is improved. In this paper, the average detection precision of the proposed model on the 2017MS COCO Test-dev dataset is 14.84% higher than that of the Mask R-CNN model, and 2.43% higher than that of the RMPE model. The frame frequency is 22 frame·s^{-1}.

Key words image processing; multi-person pose estimation; spatial transformer network; semantic information; pose distance
OCIS codes 100.4996; 100.5010; 100.3008

1 引 言

近年来，随着卷积神经网络(CNN)的发展，人体骨架关键点检测效果不断提升。人体骨架关键点对于描述人体姿态、预测人体行为至关重要，因此人体骨架关键点检测是诸多计算机视觉领域应用的基础，例如步态识别、虚拟现实、人体异常行为识别[1]、情感识别[2]等，但在室外多人环境下，姿态估计仍面临着肢体遮挡、不可见关键点、复杂背景以及实时性等挑战。

基于深度学习的多人姿态估计主要有直接回归坐标和通过热力图回归坐标两种方法。第一种代表方法有 2014 年提出的 DeepPose[3]，它将 CNN 引入姿态估计领域。在此基础上，2015 年，Fan 等[4]引入局部表观与整体视觉，提出了双源 CNN，为网络添加了先验知识。2016 年，Carreira 等[5]引入自上

收稿日期：2019-05-17；**修回日期**：2019-06-12；**录用日期**：2019-07-01
基金项目：国家自然科学基金(61671362)、陕西省科技厅重点研发计划(2019GY-022)
* **E-mail**：wp_xatu@163.com

而下的反馈机制，提出了迭代误差反馈模型，应用在前期错误检测的修正上。第二种代表方法有特征金字塔模型[6]和堆叠沙漏模型[7]等，由最初基于 CNN 与图模型网络[8]发展为 CNN 与树状结构图模型网络[9]，实现对整个人体关节的建模。多人姿态估计主流框架有自顶向下框架和基于部件框架，2017 年，Cao 等[10] 提出 CMU-Pose（Carnegie Mellon Univerisity-OpenPose）模型，基于部分亲和场连接人体各部件，并采用树结构结合匈牙利算法求解线性整数问题，提高了姿态估计运行速度。同年文献[11-13]提出双阶段 Mask R-CNN（Mask Region-CNN）模型并扩展到人体姿态估计领域，将关键点的位置建模为 one-hot 掩码并对每个掩码进行预测，提高了关键点检测精度。2017 年，上海交通大学 Fang 等[14] 提出 RMPE（Regional Multi-Person Pose Estimation）模型，加入空间变换网络以改善人体定位框不准确的问题，减少姿态估计器对人体检测框的依赖，提高了模型整体性能。

RMPE 模型中，当人体关键点完全遮挡或两个人高度重叠时，其关键点检测准确率有待提高，并且人体检测器出现漏检或重复检测会造成人体姿态估计失败。RMPE 采用自顶向下框架，其检测时间与检测人数呈线性关系，造成模型运行速度较慢，无法实现实时性需求。针对上述问题，在空间变换网络改善边界框定位精度基础上，本文基于 YOLOv3[15-16] 模型加入深度可分离卷积[17] 减少参数规模，以提高目标检测速度并改善模型提取目标提议区域能力。姿态估计模型中基于特征金字塔网络结合上下文语义信息，采用在线难例挖掘[18]（OHEM）算法解决困难关键点的检测问题，提高多人姿态估计准确率。

2 对称空间变换网络

空间变换网络（STN）赋予传统卷积裁剪、平移、缩放及旋转等特性，使模型具有空间不变性，能够自适应地将数据进行空间变换和对齐，用来提取一个高质量的人体区域框。空间反变换网络（SDTN）则用来将姿态估计结果反映射到原始图像坐标中，对称空间变换网络由 STN 和 SDTN 组成，其网络结构如图 1 所示，其中 θ 表示空间变换参数，λ 表示空间反变换参数，$T_\theta(G)$ 表示 2D 仿射变换函数，$T_\lambda(G)$ 表示 2D 反变换函数。数学上，空间变换指对应矩阵的仿射变换，图像的仿射变换可以表示为

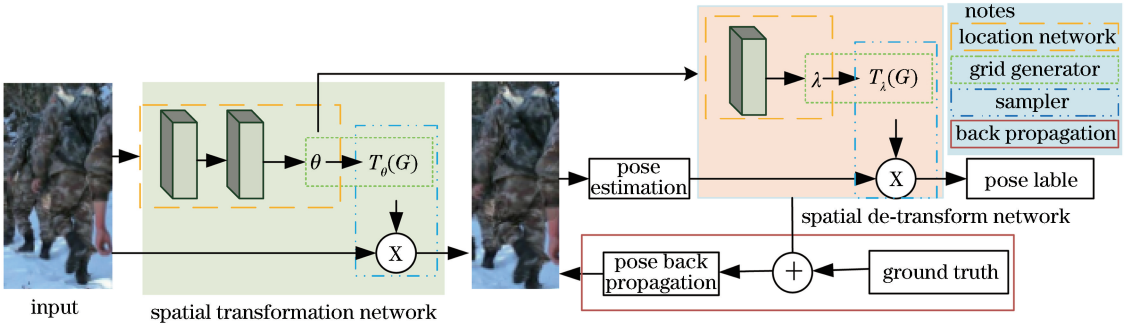

图 1 对称空间变换网络结构图

Fig. 1 Structural diagram of symmetric space transformation network

$$\begin{bmatrix} x_i^{(S)} \\ y_i^{(S)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\theta}_{11} & \boldsymbol{\theta}_{12} & \boldsymbol{\theta}_{13} \\ \boldsymbol{\theta}_{21} & \boldsymbol{\theta}_{22} & \boldsymbol{\theta}_{23} \end{bmatrix} \begin{bmatrix} x_i^{(T)} \\ y_i^{(T)} \\ 1 \end{bmatrix} = [\boldsymbol{\theta}_1 \ \boldsymbol{\theta}_2 \ \boldsymbol{\theta}_3] \begin{bmatrix} x_i^{(T)} \\ y_i^{(T)} \\ 1 \end{bmatrix}, \tag{1}$$

式中：$\boldsymbol{\theta} \rightarrow (\boldsymbol{\theta}_1 \ \boldsymbol{\theta}_2 \ \boldsymbol{\theta}_3)$ 为实域内二维空间向量；$(x_i^{(S)}, y_i^{(S)})$ 表示原图像第 i 个坐标点；S 表示转换前原图；$(x_i^{(T)}, y_i^{(T)})$ 表示仿射变换后图像第 i 个坐标点，T 表示转换后图像。首先，根据定位网络产生表示坐标映射关系的仿射变换系数 $\boldsymbol{\theta}$，计算对应输入点坐标，接着在采样器中对像素值进行填充，填充公式为

$$V_i = \sum_n \sum_m U_{nm} * k(x_i^{(S)} - m; \phi_x) * k(y_i^{(S)} - n; \phi_y), \tag{2}$$

式中：n 和 m 会遍历原始图像中的所有坐标点；U_{nm} 指原始图像通道中坐标为点 (n, m) 的像素值；V_i 为第 i 个坐标点的像素值；$k(\cdot)$ 为线性插值函数；ϕ_x 和 ϕ_y 为插值函数参数；x、y 分别表示原图像第 i 个坐标点的坐标；$*$ 表示卷积。当函数 $k(\cdot)$ 采用双线性插值时，填充公式变为

$$V_i = \sum_n \sum_m U_{nm} * \max(0, 1 - |x_i^{(S)} - m|) * \max(0, 1 - |y_i^{(S)} - n|)。 \tag{3}$$

SDTN 网络是 STN 网络的逆变换，用于计算反变换的参数 λ 可通过（1）式中参数 $\boldsymbol{\theta}\rightarrow(\boldsymbol{\theta}_1\ \boldsymbol{\theta}_2\ \boldsymbol{\theta}_3)$ 求得，其表达式为

$$\begin{bmatrix} x_i^{(\mathrm{T})} \\ y_i^{(\mathrm{T})} \end{bmatrix}=[\boldsymbol{\lambda}_1\ \boldsymbol{\lambda}_2\ \boldsymbol{\lambda}_3]\begin{bmatrix} x_i^{(\mathrm{S})} \\ y_i^{(\mathrm{S})} \\ 1 \end{bmatrix},\tag{4}$$

式中：$\boldsymbol{\lambda}\rightarrow(\boldsymbol{\lambda}_1\ \boldsymbol{\lambda}_2\ \boldsymbol{\lambda}_3)$ 是实域内二维空间向量；$(x_i^{(\mathrm{S})},y_i^{(\mathrm{S})})$ 和 $(x_i^{(\mathrm{T})},y_i^{(\mathrm{T})})$ 分别表示原图像像素点和反变换后图像像素点。$\boldsymbol{\lambda}_1,\boldsymbol{\lambda}_2,\boldsymbol{\lambda}_3$ 满足

$$[\boldsymbol{\lambda}_1\ \boldsymbol{\lambda}_2]=[\boldsymbol{\theta}_1\ \boldsymbol{\theta}_2]^{-1},\tag{5}$$

$$\boldsymbol{\lambda}_3=-1\times[\boldsymbol{\lambda}_1\ \boldsymbol{\lambda}_2]\boldsymbol{\theta}_3。\tag{6}$$

为了在 SDTN 网络中进行反向传播，$\dfrac{\partial J(\boldsymbol{W},\boldsymbol{b})}{\partial\boldsymbol{\theta}}$ 可以分解为

$$\frac{\partial J(\boldsymbol{W},\boldsymbol{b})}{\partial[\theta_1\theta_2]}=\frac{\partial J(\boldsymbol{W},\boldsymbol{b})}{\partial[\boldsymbol{\lambda}_1\ \boldsymbol{\lambda}_2]}\times\frac{\partial[\boldsymbol{\lambda}_1\ \boldsymbol{\lambda}_2]}{\partial[\boldsymbol{\theta}_1\ \boldsymbol{\theta}_2]}+\frac{\partial J(\boldsymbol{W},\boldsymbol{b})}{\partial\boldsymbol{\lambda}_3}\times\frac{\partial\boldsymbol{\lambda}_3}{\partial[\boldsymbol{\lambda}_1\ \boldsymbol{\lambda}_2]}\times\frac{\partial[\boldsymbol{\lambda}_1\ \boldsymbol{\lambda}_2]}{\partial[\boldsymbol{\theta}_1\ \boldsymbol{\theta}_2]},\tag{7}$$

$$\frac{\partial J(\boldsymbol{W},\boldsymbol{b})}{\partial\boldsymbol{\theta}_3}=\frac{\partial J(\boldsymbol{W},\boldsymbol{b})}{\partial\boldsymbol{\lambda}_3}\times\frac{\partial\boldsymbol{\lambda}_3}{\partial\boldsymbol{\theta}_3},\tag{8}$$

式中：$J(\boldsymbol{W},\boldsymbol{b})$ 为对称 STN 模型代价函数；$\boldsymbol{W}$ 和 $\boldsymbol{b}$ 均为参数矩阵。

3 多人姿态估计方法

本文提出的实时多人姿态估计模型是基于自顶向下框架的，其整体结构如图 2 所示。模型主要包括人体检测器、STN 网络与 SDTN 网络、姿态估计网络与姿态反向传播网络，以及姿态非极大值抑制网络（Pose NMS）四个部分。本文主要工作体现在以下几个方面：首先，自顶向下框架中人体检测器多采用双阶段目标检测模型，运行速度较慢，本文采用单阶段 YOLOv3 目标检测模型，有效提高了运行速度及模型泛化性，并加入可减少参数规模的深度可分离卷积，进一步提高了目标检测速度，改善了模型提取目标提议区域的能力；其次，本文姿态估计模型中基于特征金字塔网络结合上下文语义信息，采用 OHEM 算法解决困难关键点检测问题，提高了多人姿态估计准确率；最后，采用欧氏距离计算关键点之间的空间距离，判断帧内姿态相似度，剔除冗余姿态，其中地面实况数据为 2017MS COCO 数据集中人体关键点标注数据。

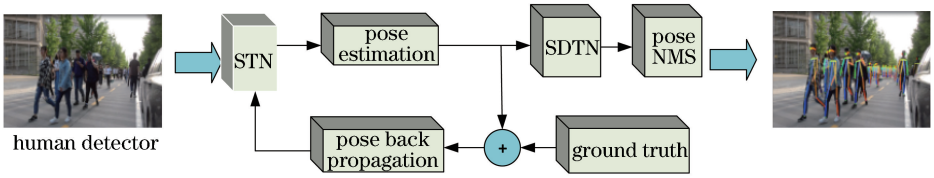

图 2 实时多人姿态估计模型

Fig. 2 Real-time multi-person pose estimation model

3.1 空间可分离卷积

YOLOv3 模型借鉴特征金字塔网络，采用逻辑回归代替 Softmax 函数作为分类器，目标检测速度大幅度提升。本文基于 YOLOv3 目标检测模型，在图像卷积过程中，加入深度可分离卷积，能够有效减少模型的参数规模，提高目标检测速度。

标准卷积中，每个输入通道均与一个特定卷积核进行卷积，将来自所有通道的卷积结果的总和作为最终结果。深度可分离卷积中，首先进行深度卷积，分别对每个输入通道执行卷积，然后逐点进行卷积。与标准卷积相比，这种卷积结构可以极大地减小网络模型的参数数量和计算量，并且不会造成明显的精确度损失。例如传统卷积核是对 3 个通道同时进行卷积，即 3 个通道在一次卷积后，输出一个卷积值。深度可分离卷积是用三个卷积核对三个通道分别进行卷积，这样在一次卷积后输出 3 个卷积值，然后再通过一个 $1\times1\times3$ 的卷积核，得到最终卷积值。随着提取属性的增加，深度可分离卷积能够节省更多参数，减少模型计算量。

当输入图像大小为 $M\times M\times N$，卷积核大小为 $K\times K\times N\times P$，步长为 1 时，标准卷积所需参数规模大小 P_{SC} 和卷积操作计算量 C_{SC} 分别为

$$P_{\mathrm{SC}}=K\times K\times N\times P,\tag{9}$$

$$C_{\mathrm{SC}}=M\times M\times K\times K\times N\times P。\tag{10}$$

深度可分离卷积所需参数规模大小 P_{DSC} 和卷积操作计算量 C_{DSC} 分别为

$$P_{\mathrm{DSC}}=K\times K\times N+N\times P,\tag{11}$$

$$C_{\mathrm{DSC}}=M\times M\times K\times K\times N+M\times M\times N\times P。\tag{12}$$

其参数规模变化 P_{C} 和减少率 P_{CR} 计算定义式为

$$P_{\mathrm{C}}=\frac{P_{\mathrm{DSC}}}{P_{\mathrm{SC}}}=\frac{1}{P}+\frac{1}{K^2},\tag{13}$$

$$P_{CR}=\frac{|P_{DSC}-P_{SC}|}{P_{SC}}=\frac{|K^2\times P-K^2-P|}{K^2+P}。\tag{14}$$

卷积操作计算量变化与减少率计算方式同(13)式、(14)式。

3.2 姿态估计网络

本文根据人体关键点检测的不同难易程度再结合在线难例挖掘算法对关键点进行检测。首先基于特征金字塔网络定位如头、肩膀、手肘等易识别的关键点,再结合上下文语义信息定位如脚踝、手腕、臀部等困难的关键点,最终完成整个人体关键点检测。损失函数采用均方误差(MSE,L2Loss),网络结构如图 3 所示。

网络架构基于残差网络模型,把不同卷积特征的最后残差块分别表示为 C_2,C_3,C_4 及 C_5,并使用 3×3 卷积滤波器生成关键点的热力图。浅层特征在定位上有较高的空间分辨率,但在关键点检测上语义信息较少,深层特征如 C_4、C_5 语义信息较多,但空间分辨率较低,因此经常引入 U 型结构同时保留特征层的空间分辨率和语义信息。

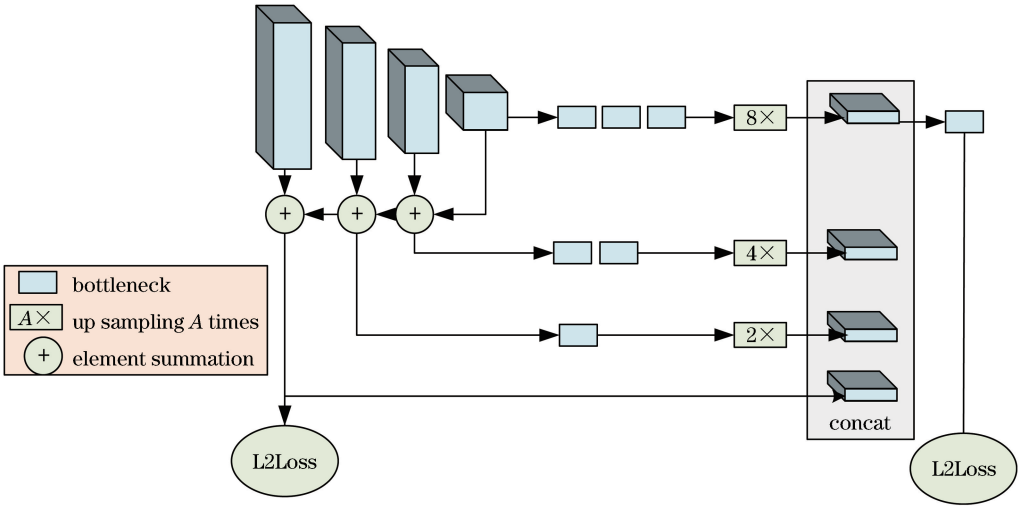

图 3 姿态估计网络模型

Fig. 3 Pose estimation network model

本文在特征金字塔的基础上完善了 U 型结构,在上采样过程中,逐像素相加求和前使用 1×1 卷积核,并在姿态识别后端网络加入更多的 Bottleneck 模块来处理更深的特征,使其在较小的空间尺度实现效率和性能的权衡。随着训练的进行,网络会倾向于关注比例较多的易识别关键点,对于识别难度较大的关键点,比如遮挡等情况,网络的关注会逐渐降低,因此本文根据训练损失在线选择困难关键点,并只从已选择的关键点反向传播梯度,平衡网络对难易程度不同关键点的关注。

目标检测沿用图像分类的分类思想,但图像分类的数据集和目标检测的数据集存在天然的差距,这导致目标检测的目标框和背景框之间存在严重的不平衡。困难负样本挖掘(HNM)算法被用来解决这个问题,其关键思想是逐渐增加错误检测的样本数量。但由于此算法需要迭代交替训练,用样本集更新模型,无法实现在线优化算法,因此本文基于 OHEM 算法将迭代交替训练步骤与随机梯度下降(SGD)算法结合起来,实现在线困难关键点选择。OHEM 算法核心是选择一些困难样本作为训练样本从而改善网络参数效果,首先通过前向传播算法计算所有关键点的热力图损失值,然后根据排序选择前 K 个损失值最大的样本进行反向传播,更新模型的权重。本文基于像素层次计算姿态估计网络输出的热力图中的损失值,损失函数采用最小平方误差损失函数,模型参数梯度更新计算公式为

$$\boldsymbol{\theta}=\boldsymbol{\theta}-\eta\cdot\nabla_{\boldsymbol{\theta}}J(\boldsymbol{\theta};x^{(i)};y^{(i)}),\tag{15}$$

式中:$x^{(i)}$ 为训练样本;$y^{(i)}$ 为样本标签;$\boldsymbol{\theta}$ 为给定待优化参数;$J(\cdot)$为目标函数;η 为学习率,决定每一时刻的更新步长;算法通过沿梯度$\nabla_{\boldsymbol{\theta}}J(\cdot)$的相反方向更新 $\boldsymbol{\theta}$ 来最小化 $J(\cdot)$。加入在线困难关键点挖掘算法,能够有效解决关键点遮挡、不可视等问题,提高了脚踝、臀部等困难关键点的检测精度。

3.3 姿态非极大值抑制

人体检测器不可避免地会产生冗余检测,导致姿态估计网络生成了冗余姿态,因此,本文提出姿态非极大值抑制模型来剔除冗余姿态。在非极大值抑制算法基础上,本文首先选取置信度最大的姿态作

为参考，然后重新定义消去原则去除靠近该区域的区域框，重复此步骤直至每一个检测框都是唯一出现。

根据姿态距离自定义函数 $D\_\mathrm{Sim}(P_1,P_2|\Lambda)$ 来衡量同一帧中姿态 P_1 和 P_2 之间的相似度，定义 γ 为消除原则的阈值，Λ 表示函数 $D(\cdot)$ 的一个参数集合，消除原则表示为

$$f(P_1,P_2\mid\Lambda,\gamma)=1[D\_\mathrm{Sim}(P_1,P_2\mid\Lambda,\lambda)\leqslant\gamma]。\tag{16}$$

如果 $D\_\mathrm{Sim}(\cdot)$ 小于阈值 γ，那么 $f(\cdot)$ 的输出是1，表示姿态 P_i 应该被消除，因为对于参考的 P_2 来说 P_1 是冗余的。$P_1^{(n)}$ 和 $P_2^{(n)}$ 代表姿态 P_1 和 P_2 的第 n 个关键点，姿态 $P_1^{(n)}$ 和 $P_2^{(n)}$ 的中心检测框分别为 $\mathrm{Box}(P_1^{(n)})$ 和 $\mathrm{Box}(P_2^{(n)})$，$c_1^{(n)}$ 和 $c_2^{(n)}$ 分别为 $P_1^{(n)}$ 和 $P_2^{(n)}$ 的置信度得分，衡量帧内两个姿态相似度的函数为

$$S\_\mathrm{Sim}(P_1,P_2\mid\sigma_1)=\begin{cases}\sum_n\tanh\dfrac{c_1^{(n)}}{\sigma_1}\cdot\tanh\dfrac{c_2^{(n)}}{\sigma_1},P_2^{(n)}\subset\mathrm{Box}(P_1^{(n)})\\0,\ \mathrm{otherwise}\end{cases},\tag{17}$$

tanh 函数可以滤掉低置信度的姿态，当两个姿态的置信度都比较高的时候，上述函数的输出接近1，此距离表示了姿态间不同部位的匹配数。将关键点之间空间相似度定义为

$$R\_\mathrm{Sim}(P_1,P_2\mid\sigma_2)=\sum_n\exp\left[-\frac{\sqrt{(P_1^{(n)}-P_2^{(n)})^2}}{\sigma_2}\right]。\tag{18}$$

本文采用欧氏距离计算关键点之间的空间距离，最终姿态 P_1 和 P_2 之间的距离公式为

$$D\_\mathrm{Sim}(P_1,P_2\mid\Lambda)=S\_\mathrm{Sim}(P_1,P_2\mid\sigma_1)+\beta R\_\mathrm{Sim}(P_1,P_2\mid\sigma_2),\tag{19}$$

式中：β 是一个权重系数，用来平衡这两种距离；Λ 表示 $\{\sigma_1,\sigma_2,\beta\}$ 参数集合，σ_1 和 σ_2 由数据驱动的方式获得。

4 分析与讨论

本文实验环境如下：操作系统为 Ubuntu16.04，CPU 环境为 2x Intel(R)Xeon(R)Gold6128CPU@3.40 GHz，内存 32G，1T 7200 SATA3.5＋512 G SSD，GPU 环境为 2x NVIDIA Quadro P2000 5 GB 显卡。多人姿态估计模型基于 Pytorch0.4.0＋Python3.6.2 建立。为了更好展示本文所设计模型性能，在相同硬件环境下分析对比了当前流行的三种姿态估计模型：Mask R-CNN 模型、CMU-Pose 模型以及 RMPE 模型，其环境配置如表1所示。

表1 算法模型及环境配置

Table 1 Algorithm model and environment configuration

Model	Framework	Programming language
CMU-Pose[10]	Caffe	Python3.6.2
Mask R-CNN[11]	TensorFlow1.3.0＋Keras2.2.6	Python3.6.2
RMPE[14]	Pytorch0.4.0	Python3.6.2
Proposed model	Pytorch0.4.0	Python3.6.2

4.1 定性分析

本文以校园采集图像、网络图像及 2017COCO 数据集图像共四张典型图像为例进行多人姿态估计，展示尺度变化、密集人群、遮挡和复杂姿态四种场景下模型的泛化能力。模型姿态估计效果如图4所示。

图4(a)、(e)、(i)、(m)为 Mask R-CNN 模型姿态估计结果，图4(b)、(f)、(j)、(n)为 CMU-Pose 模型姿态估计结果，图4(c)、(g)、(k)、(o)为 RMPE 模型姿态估计结果，图4(d)、(h)、(l)、(p)本文设计模型姿态估计结果。

在尺度变化场景中，人体尺度较小时[图4(a)]未能正确检测到人体关键点，人体腿部及胳膊处关键点检测均出现错误。由于图像最中间五个目标紧密连接，只成功检测到最左边边上人体关键点，且人体目标脚踝关键点检测错误，如图4所示。在目标距离较近时姿态估计出现错误，如图4(c)所示。

在密集场景中，图4(e)只检测到最后一排人体的上半身姿态，图4(f)将图像左右两侧人体的肩膀处关键点错误连接。图4(g)无法检测最后一排被遮挡的头部姿态，且手部关键点检测准确率较低。

在肢体遮挡场景中，中间目标遮挡了后面目标，导致对图4(i)遮挡目标姿态估计错误。图像前后目标也存在尺度变化较大问题，导致未能检测到图4(j)后面背景目标。图4(k)目标运动导致存在虚影且人体左臂与上半身肢体重叠，所以未能检测到左肩、左臂手肘以及手部关键点。

在人体复杂姿态场景中，图4(m)姿态估计失败，由于复杂姿态导致人体上半身遮挡较多，因此图4(n)上半身姿态估计失败，图4(o)同样在肢体重叠较多区域姿态估计失败。

实验表明，在上述四种典型场景下，当前流行的三种姿态估计模型处理效果有待提高。由于本文将简单关键点的特征信息共享至网络深层特征，结合

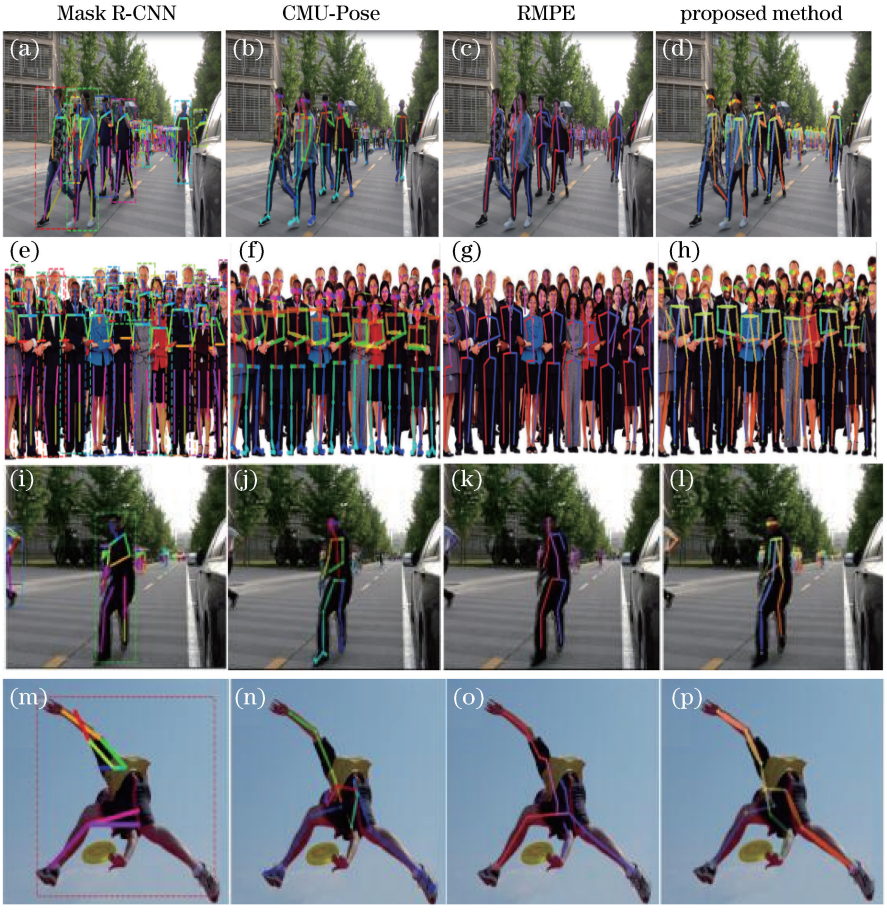

图 4 不同场景下各模型结果对比。(a)～(d)尺度变化。(e)～(h)密集人群;(i)～(l)遮挡;(m)～(p)复杂姿态

Fig. 4 Comparison of results in different scenarios for each model. (a)-(d) Scale change; (e)-(h) dense population; (i)-(l) occlusion; (m)-(p) complex pose

上下文语义信息检测困难关键点,并加入 OHEM 算法加强模型对困难关键点的关注度,有助于提高困难关键点检测精度,因此,本文设计模型均成功实现姿态估计,各个关键点检测效果明显优于另外三种模型。

4.2 定量分析

本文空间变换网络中采用 ResNet-18 作为定位网络,姿态估计网络是在特征金字塔网络基础上构建的。在网络设计中加入多个 Bottleneck 模块,融合不同层特征,加入上下文语义信息实现困难关键点检测。在模型训练中,图片裁剪宽高比为 384∶288,采用随机翻转策略将图片随机旋转(−45°～+45°)并改变图像尺度,尺度变化为 0.7、1、1.35 三个不同尺度。模型训练数据集为 2017MS COCO 数据集,包括 57×10^3 图像和 150×10^3 人体实例,训练过程中采用 Adam 算法,迭代更新网络权重,每 360 万次迭代后将学习率降低 1/2,初始学习率为 5×10^{-4}。训练过程中用 Pytorch 框架中 nn. Functional. interpolate(·)函数代替 nn. Upsampling(·)函数进行上采样操作。

MS COCO 评估指标中,对象关键点相似性(OKS)定义式为

$$R_{\mathrm{OKS}}=\frac{\sum_i \exp\{-d_{pi}^2/2S_p^2\sigma_i^2\}\delta(v_{pi}=1)}{\sum_i \delta(v_{pi}=1)},\tag{20}$$

式中:p 为地面实况中人的 id;i 表示关键点的 id;d_{pi} 表示地面实况中每个人关键点与预测关键点的欧氏距离;S_p 表示当前人的尺度因子,即此人在地面实况中所占面积的平方根;σ_i 表示第 i 个关键点的归一化因子;v_{pi} 代表第 p 个人的第 i 个关键点是否可见;δ 为将可见点选出来进行计算的函数。AP 即所有 10 个 OKS 阈值的平均精确率,AR 即所有 10 个 OKS 阈值的平均召回率。$\mathrm{AP}_{@0.5}$ 表示 OKS 为 0.5 时 AP 值,$\mathrm{AP}_{@0.75}$ 表示 OKS 为 0.75 时的 AP 值,AP_{m} 表示中等目标 AP 值,面积大小范围为(322,962),AP_{l} 表示大目标 AP 值,面积大小范围

为(962,+∞),AR 参数含义同 AP。

本文设计模型与目前领先的姿态估计模型性能在 2017COCO Test-dev 数据集的对比如表 2 所示。

表 2 各姿态估计模型性能对比

Table 2 Comparison of performance of each pose estimation model

Model	AP	$AP_{@0.5}$	$AP_{@0.75}$	AP_m	AP_l
CMU-Pose	61.8	84.9	67.5	57.1	68.2
Mask R-CNN	63.1	87.3	68.7	57.8	71.4
RMPE	72.3	89.2	79.1	68.0	78.6
Proposed model	74.1	92.5	80.5	70.6	79.5

本文数据预处理中采用随机翻转与旋转策略,且生成三个不同的图像尺度,既扩增了数据规模,也使得数据具有良好的尺度不变性和旋转不变性。网络结构设计中融合不同层目标特征,加入上下文语义信息,有利于提高困难关键点的检测准确率。由 2107 COCO Test-dev 数据集评估指标可得,本文姿态估计模型关键点检测精度得到提高,平均检测精度比 Mask R-CNN 模型提升了 14.84%,比 RMPE 模型提升了 2.43%.

1)人体检测器参数规模分析

本文分析对比了所提模型与 YOLOv3 目标检测模型的参数规模,其数据见表 3。

表 3 各人体检测算法参数规模对比

Table 3 Comparison of parameters of each human detection algorithm

Model	Data set	Running speed / (frame · s^{-1})	Parameter size /MB	Calculated amount /10^9
YOLOv3[15]	MS COCO	51	237	65.86
Proposed model	MS COCO	64	195	44.32

由表 3 可知,经过深度可分离卷积操作后,参数规模约减少了 17.72%,计算量约减少了 32.71%,帧频为 64 frame/s。

2)姿态估计模型 P-R 性能分析

本文模型在输入图像尺寸为 256 pixel×192 pixel和 384 pixel×288 pixel 时不同交并比下 AP 与 AR 值如表 4 所示。

由表 4 可知,随着输入图像尺寸的增加,本文姿态估计模型的的平均准确率有所提高,模型的性能也随之提升,其原因在于图像尺寸和分辨率越大,网络提取的特征图分辨率就越高,模型的表现也就越好。

表 4 不同输入下模型 AP-AR 值

Table 4 AP-AR values of model under different inputs

Input	AP	$AP_{@0.5}$	$AP_{@0.75}$	AP_m	AP_l	AR	$AR_{@0.5}$	$AR_{@0.75}$	AR_m	AR_l
256 pixel×192 pixel	71.2	91.4	78.3	68.5	75.2	74.3	92.2	80.9	71.3	78.9
384 pixel×288 pixel	74.1	92.5	80.5	70.6	79.5	76.8	93.2	82.5	73.0	82.6

5 结 论

本文基于自顶向下框架,设计了一套实时多人姿态估计模型,通过将深度可分离卷积算法加入单阶段 YOLOv3 目标检测算法中,有效提高了多人姿态估计模型关键点检测速度;基于特征金字塔网络结合上下文语义信息,采用在线难例挖掘算法有效提高了关键点检测准确率。在 2017COCO Test-dev 数据集上,本文设计模型关键点检测精度相对 Mask R-CNN 提升了 14.84%,比采用 Faster-RCNN 作为人体检测器的 RMPE 模型提升了 2.43%。在本文实验环境下,多人姿态估计网络模型实时运行速度达到 22 frame/s。在未来,可以深入研究自顶向下框架与基于部件框架的结合,设计一套不依赖于人体检测器且关键点检测准确率极高的端到端多人姿态估计模型。

参 考 文 献

[1] Yan S, Xiong Y, Lin D. Spatial temporal graph convolutional networks for skeleton-based action recognition[C]// Thirty-Second AAAI Conference on Artificial Intelligence, February 2-7, 2018, Hilton New Orleans Riverside, New Orleans, Louisiana, USA. USA: AAAI, 2018: 7444-7452.

[2] Jiang M X, Hu M, Wang X H, et al. Dual-modal emotion recognition based on facial expression and body posture in video sequences[J]. Laser & Optoelectronics Progress, 2018, 55(7): 071004.
姜明星, 胡敏, 王晓华, 等. 视频序列中表情和姿态的双模态情感识别[J]. 激光与光电子学进展, 2018, 55(7): 071004.

[3] Toshev A, Szegedy C. DeepPose: human pose estimation via deep neural networks[C]// 2014 IEEE Conference on Computer Vision and Pattern

Recognition, June 23-28, 2014, Columbus, OH, USA. New York: IEEE, 2014: 1653-1660.

[4] Fan X C, Zheng K, Lin Y W, et al. Combining local appearance and holistic view: dual-source deep neural networks for human pose estimation[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE, 2015: 1347-1355.

[5] Carreira J, Agrawal P, Fragkiadaki K, et al. Human pose estimation with iterative error feedback[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 4733-4742.

[6] Yang W, Li S, Ouyang W L, et al. Learning feature pyramids for human pose estimation[C]// 2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE, 2017: 1290-1299.

[7] Newell A, Yang K Y, Deng J. Stacked hourglass networks for human pose estimation[M]// Leibe B, Matas J, Sebe N, et al. Computer vision-ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016, 9912: 483-499.

[8] Tompson J J, Jain A, LeCun Y, et al. Joint training of a convolutional network and a graphical model for human pose estimation[C]// Advances in Neural Information Processing Systems 27 (NIPS 2014), December 8-13, 2014, Montreal, Quebec, Canada. Canada: NIPS, 2014.

[9] Yang W, Ouyang W L, Li H S, et al. End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 3073-3082.

[10] Cao Z, Simon T, Wei S E, et al. Realtime multi-person 2D pose estimation using part affinity fields[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 1302-1310.

[11] He K M, Gkioxari G, Dollar P, et al. Mask R-CNN[C]// 2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE, 2017: 2980-2988.

[12] Ren S Q, He K M, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[13] Feng X Y, Mei W, Hu D S. Aerial target detection based on improved Faster R-CNN[J]. Acta Optica Sinica, 2018, 38(6): 0615004.
冯小雨, 梅卫, 胡大帅. 基于改进 Faster R-CNN 的空中目标检测[J]. 光学学报, 2018, 38(6): 0615004.

[14] Fang H S, Xie S Q, Tai Y W, et al. RMPE: regional multi-person pose estimation[C]// 2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE, 2017: 2353-2362.

[15] Redmon J, Farhadi A. YOLOv3: an incremental improvement[J/OL]. (2018-04-08)[2019-05-16]. https://arxiv.org/abs/1804.02767.

[16] Wei Y M, Quan J C, Houyu Q Y. Aerial image location of unmanned aerial vehicle based on YOLO v2[J]. Laser & Optoelectronics Progress, 2017, 54(11): 111002.
魏湧明, 全吉成, 侯宇青阳. 基于 YOLO v2 的无人机航拍图像定位研究[J]. 激光与光电子学进展, 2017, 54(11): 111002.

[17] Chen L C, Zhu Y K, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[M]// Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11211: 833-851.

[18] Shrivastava A, Gupta A, Girshick R. Training region-based object detectors with online hard example mining[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 761-769.