基于实例分割的多人姿态检测与跟踪算法的设计与实现

**信息安全 15143103 刘方瑞**

**指导教师 马伟**

摘要

提出基于实例分割的多人姿态估计算法，在单张图像上准确提取多人人体姿态，以实现序列图像多人姿态的检测与跟踪。该算法提出一个统一的深度学习网络结构，融合实例分割与多人姿态估计任务，以有效区分不同人体区域。为了推断被遮挡部分的人体姿态，设计并训练了一个人体区域注意力子模块。鉴于现有实例分割真值缺失人体被遮挡部分的数据，难以用于人体区域注意力子模块的监督学习，参考弱监督学习策略，提出使用关键点与实例分割真值共同构建弱监督信息训练注意力子模块。在提升多人姿态估计准确度的同时，所提出网络能够借助姿态线索反向优化实例分割结果。通过对比实验验证了所提出算法的有效性。

**关键词：**计算机视觉；人体姿态估计；实例分割；注意力机制；弱监督学习

Abstract

We propose a novel deep learning architecture for multi-person pose estimation in single images, to achieve multi-person pose detection and tracking in image sequences. It integrates the tasks of instance segmentation and multi-person pose estimation in a unified framework to distinguish different persons. In order to infer poses in occluded human parts, we design a human-region spatial attention module and train it with indirect constraints from tasks of pose key-points and instance segmentation, as the occluded body parts are incomplete in the ground truth of instance segmentation annotations. Beside of improvement on key-points accuracy, the proposed network obtains better instance segmentation by referring to cues from pose estimation. We validate the proposed architecture via comparison experiments.

**Keywords:** Computer Vision; Human Pose Estimation; Instance Segmentation; Attention Mechanism; Weakly Supervised Learning

**1. 绪论**

人体姿态检测与跟踪旨在从传感器数据中捕捉多人姿态信息，在各领域具有广阔的应用前景。相比基于组合传感器的姿态采集与追踪系统，基于单目相机的多人姿态估计方法不受限于室内场景，因此拥有更高的研究与应用价值。然而，基于单目相机的多人姿态提取面临诸多技术难点，其中之一是多人之间的互遮挡问题。现有方法难以根据可见人体部分推断不可见部位的姿态。鉴于上述事实，本文在研究现有方法的基础上，提出基于实例分割的多人姿态估计方法，以更好地解决多人之间的互遮挡问题。

**2. 基于实例分割的多人姿态估计方法**

所提出网络包含检测和双任务融合优化两个模块。检测模块计算姿态关键点位置和实例分割的初始结果。融合优化模块接收来自检测模块的特征以及初始估计值，联合优化人体姿态关键点位置和实例分割结果。特别地，网络将分割任务加入到多阶段迭代优化的框架内，并采取了多阶段堆叠的优化结构和中继监督策略共同优化关键点与实例分割结果。同时，本文引入人体注意力机制，在复合损失函数与单通道结构的间接约束下，鼓励其生成被遮挡的人体区域，并排除周边关键点检测对特定人体的影响。

2.1 检测模块

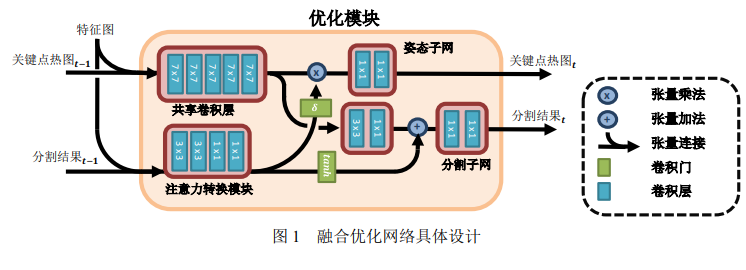
检测模块包含特征提取网络与检测回归网络。在特征提取网络中，结合残差网络[1]与特征金字塔[2]结构提取图像中各个尺度的特征。之后，检测回归网络根据特征提取网络提供的特征图，粗略预测关键点位置和实例分割结果。

2.2 融合优化网络

本文设计了多个具有相同结构的优化模块同时优化实例分割与姿态估计。所提出的优化模块通过引入软注意力，以强化特定人体空间区域的特征响应，从而改善网络在遮挡情况下对人体姿态信息的提取能力。同时软注意力可增强实例分割的特征表达，在模块间传递并优化分割信息。

（1）优化模块结构设计

优化模块结构设计如图1所示。该模块首先将多任务分支的信息从当前阶段传递到下一阶段。同时，融合两任务之间的特征，使网络结合来自不同任务的线索以得到更准确的结果。网络结构在每个分支都接收来自上一阶段的中间估计结果，并在当前阶段进一步提升结果准确度。两个任务特征通过人体空间注意力相互融合以达到交叉优化的目标。



（2）弱监督与注意力机制

本文使用了弱监督方法约束空间注意力的生成。空间注意力的目标在于获得特定人体的完整区域。实例分割的标注真值未能将被遮挡的人体区域考虑在内，而部分姿态估计真值给出了被遮挡躯干的关键点位置。本文参考了弱监督策略下的学习过程，通过构建间接作用于单通道的空间注意力的损失函数约束，在不完整的标注信息下监督空间注意力的生成。姿态估计信息能够为空间注意力提供结构化的线索，同时实例分割为注意力提供清晰地可见部分的边界监督信息。本文以空间注意力为桥梁设计网络结构，将两任务分支交叉连接，在复合损失函数与单通道结构的共同作用下，间接约束网络生成考虑人体遮挡区域的空间注意力，在空间上选择并强化两任务中对应位置的特征，从而增强多人姿态估计的结果。

2.3 网络训练

本文监督网络的损失函数可以被分为三部分：检测损失函数、关键点分支损失函数与实例分割分支损失函数（具体定义见学位论文）。为了平衡每个分支的收敛速率，本文设计了β和α分别控制目标检测与实例分割的损失函数权重（具体定义见学位论文）。本文提出的优化模块结构涉及到两支子网相互交叉与传递，为了证明网络能够在添加结构的条件下保证两支网络的正常收敛（具体证明见学位论文），给出了损失函数中不同项对于梯度的影响，从而解释了弱监督信息能使用复合损失函数监督注意力的生成。

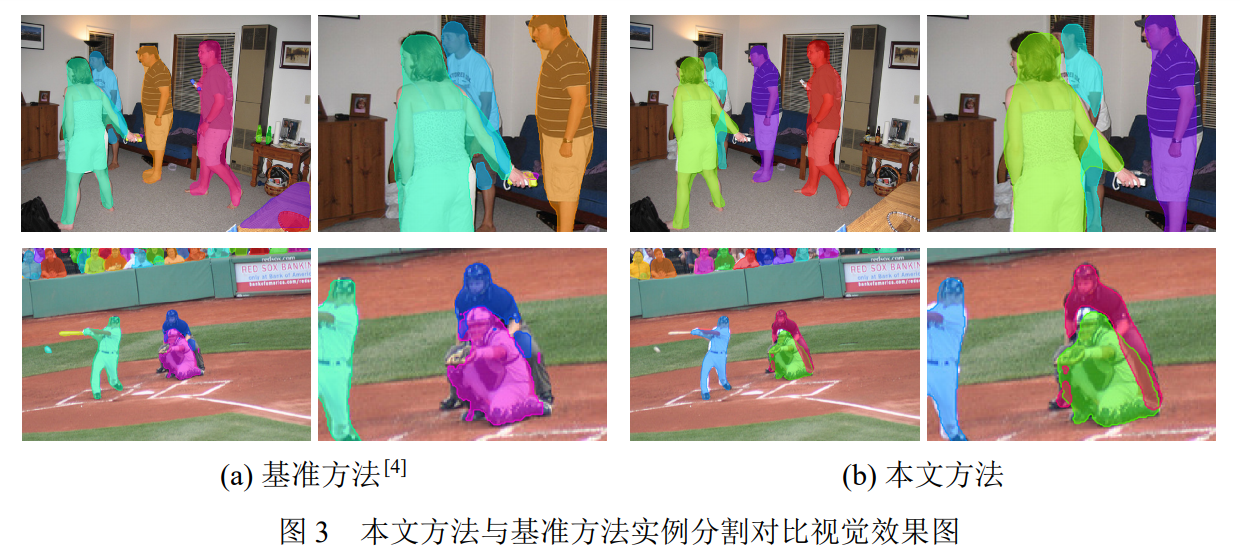
**3. 实验验证与分析**

本文在COCO2017数据集上对比了现有方法与本文方法的性能指标，并通过多个对比消融实验证明了本文中提出的核心模块的有效性。

3.1 视觉效果与分析

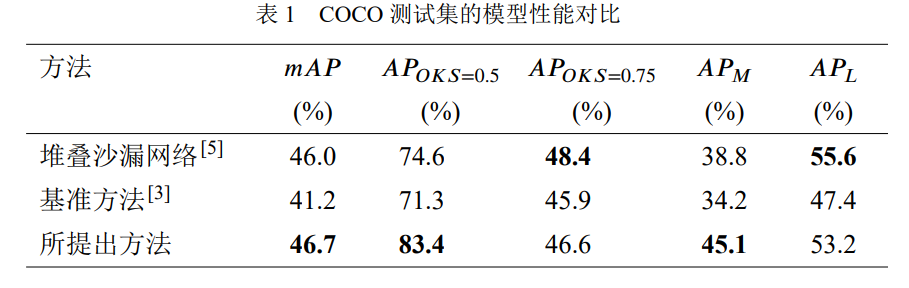


如图2所示，本文方法可以比基准方法更好地关注单个实例。通过自生成空间注意力约束，本文方法能够给出更加规整的关键点预测结果，并将结果限制于有限的特定人体区域内。基准方法中给出的多人姿态估计会受到相邻实例的干扰，影响关键点结果的完整性。此外，对于一些完全遮挡下的人体区域，本文方法可以通过注意力生成的关注区域引导姿态分支给出姿态估计。如图2第一行基准方法与本文方法可视化结果对比所示，在注意力的约束下所提出的方法能够排除临近人体躯干的干扰，给出更整齐准确的姿态估计结果。



如图3所示，实例分割结果相比于基准方法能够更多地考虑人体全局的特征。在第一行中被手臂遮挡的人体部位在基准方法给出的结果中是上下分离的，并且还出现下半身缺失分割的情况，而在本文方法中，不仅给出了大致结果，同时还连接了一部分断开的身体部分。如图3中的第二行所示，本文方法给出的实例分割结果对遮挡情况的应对相比基准方法得到了明显的改善。

3.2 性能对比与分析



如表1所示，与堆叠沙漏网络相比使用本文提出的优化模块能够显著改善关键点在APOKS=0.5的得分(7.5AP)。这部分的提升主要是来自空间注意力提供的对人体的精确划分，使得给出的姿态估计更加完整。同时，本文方法在对小目标的姿态检测结果性能APM上性能较堆叠沙漏网络提高了4.2AP。而之后在此基础上额外增加优化模块的本文方法+在原有策略基础上提升0.8mAP，在APOKS=0.5与APM指标上比沙漏网络分别高出8.8AP与6.3AP。整体而言，本文方法在AP指标上比现有的自顶向下的堆叠沙漏网络高出0.7mAP，并能够给出更加完整的姿态估计结果。

**4. 结论**

本文提出一种全新的融合人体实例分割信息的多人姿态检测网络结构。该结构将实例分割信息加入到多阶段的优化网络设计中，同时优化分割结果与姿态估计结果。通过大量定量与定性实验展示网络模型同时给出分割结果和姿态估计结果的性能，并证明本文提出弱监督下空间注意力的有效性。

参考文献

[1] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-778.

[2] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2017: 2117-2125.

[3] Wei S E, Ramakrishna V, Kanade T, et al. Convolutional pose machines[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2016: 4724-4732

[4] He K, Gkioxari G, Dollár P, et al. Mask R-CNN[C]//IEEE International Conference on Computer Vision. 2017: 2961-2969.

[5] Newell A, Yang K, Deng J. Stacked hourglass networks for human pose estimation[C]//European Conference on Computer Vision. 2016: 483-499.