计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(11):138−144 [doi: 10.15888/j.cnki.csa.008128]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 RetinaNet-CPN 网络的视频人体关键点检测①

包晓安, 吉鹏飞

(浙江理工大学, 杭州 310018)
通讯作者: 吉鹏飞, E-mail: 18861120983@163.com

摘　要: 为解决基于视频流的人体关键点检测效果不佳及视频流切片后可能会发生运动模糊的问题, 提出了一种改进的 RetinaNet-CPN 网络对人体关键点进行检测, 有效解决切片后运动模糊图像的干扰并提高了人体关键点的检测准确率. 视频流切片后, 先用改进的 RetinaNet 网络检测出图片中的所有人并对每个目标框做模糊检测, 对大于阈值的目标框做去模糊处理, 最后用引入注意力机制的 CPN 网络提取关键点. 将 RetinaNet 衡量预测框与真实框差异的 IOU 函数改成 DIOU 后, 在仿真实验中目标检测 AP 提高了近 3%; 对于模糊的图片, 利用匀速直线运动频谱特征估算出的模糊核与实际模糊核相差不大, 对其做去模糊处理后基本能恢复出原清晰图片; 同时引入注意力机制为各通道和特征层分配合理的权重, 使得 CPN 检测 AP 提高近 1%, AR 提升 0.5%.

关键词: 人体关键点检测; RetinaNet 网络; CPN 网络; 模糊检测; 频谱特征; 注意力机制

引用格式: 包晓安,吉鹏飞.基于 RetinaNet-CPN 网络的视频人体关键点检测.计算机系统应用,2021,30(11):138–144. http://www.c-s-a.org.cn/1003-3254/8128.html

Video Human Body Keypoint Detection Based on RetinaNet-CPN Network

BAO Xiao-An, JI Peng-Fei

(Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract: Concerning the problem of poor detection of human body keypoints based on video streams and possible motion blur after video stream slicing, an improved RetinaNet-CPN network is proposed to detect the keypoints, avoiding the interference of motion-blurred images after slicing and improving the detection accuracy of the keypoints. After the video stream is sliced, the improved RetinaNet network is first used to detect all the people in the picture and perform fuzzy detection on each target frame. The target frame larger than the threshold is deblurred, and finally, the keypoints are extracted with the CPN network with the attention mechanism. After the IOU function of RetinaNet to measure the difference between the predicted frame and the real frame is changed into DIOU, the target detection AP increases by nearly 3% in the simulation experiment. For blurry pictures, the blur kernel estimated with the spectrum feature of uniform linear motion is slightly different from the actual blur kernel, and the original clear picture can be restored after the deblurring. At the same time, the attention mechanism is adopted to assign reasonable weights to each channel and feature layer, which increases the CPN detection AP by nearly 1% and the AR by 0.5%.

Key words: human body keypoint detection; RetinaNet network; CPN network; fuzzy detection; spectrum feature; attention mechanism

① 基金项目: 浙江省自然科学基金 (LQ20F050010)
Foundation item: Natural Science Foundation of Zhejiang Province (LQ20F050010)
收稿时间: 2021-01-13; 修改时间: 2021-02-07; 采用时间: 2021-02-23; csa 在线出版时间: 2021-10-22

人体关键点检测技术在姿态估计中具有重要意义, 并且在安防, 游戏娱乐等行业应用前景广阔. 多人关键点检测主要有自顶向下和自底向上这两种策略, 自顶向下的策略是先将图片中目标分割后再分别检测每个目标的关键点; 而后者是先检测出所有关键点然后再用特定的算法将这些关键点与每个目标进行匹配. 2014 年 Toshev 等首次提出基于深度学习的 deeppose 模型[1] 来进行关键点检测. 之后为提高检测效率, Tompson 等[2] 提出用 heatmap 回归关键点. Carreira 等提出 CPM[3], 并引入中间监督机制来防止梯度消失, 能够很好地融合图片的各种信息, 但它参数较多, 检测速度较慢. Newell 等提出了 Hourglass module 结构[4], 它能够对一些难检测的关键点做出预测, 但效果不佳. CMU 提出一种基于 Bottom-up 策略的 Openpose[5], 它检测出所有关键点后, 通过二分图最大权匹配算法来对关键点进行拼接, 得到每个人的骨架. 2017 年旷视提出了 CPN 网络[6], 它能够对一些遮挡的, 难以检测的点做出很好的预测. 2019 年微软提出了 HRNet[7], 它将不同大小的特征层由串行变成了并行, 减小了上采样下采样的信息的丢失.

对于多人关键点检测, 两种策略各有优劣, 使用自底向上的方法虽然计算量较小, 但当目标密集, 该算法容易将关键点误判, 并且如果目标尺度较小可能会漏检; 自顶向下方法准确率更高, 但检测结果受目标检测的影响较大, 并且计算量与标记框数量成正比. 为了避免错判和漏检, 还是用自顶向下的方案更合适, 因此目标检测标记框的准确率及关键点检测前的图片质量就显得尤为重要. 常见的单阶段目标检测算法有 YOLO、SSD[8], RetinaNet[9] 等; 常见的双阶段算法有 Faster RCNN、Mask RCNN[10] 以及 CascadeRCNN[11] 等. 与单阶段相比双阶段要先用 RPN 等算法筛选出一定数量的候选框而后再分类和回归, 准确率高但速度比较慢.

目标检测网络训练时通常采用 IOU 来划分正负样本, 但它不能够很好地衡量预测框和实际框的重叠程度, 并且当重叠度为 0 时, 它的损失函数值也为 0, 此时没有梯度返回无法学习; 另一方面由于关键点检测网络没有给每个通道和特征层分配合适的权重, 其精度仍有提升空间. 同时在实际应用中往往是基于视频流切片的, 难免会发生运动模糊现象影响图片质量进而降低关键点检测的准确率. 为此本文提出了一种改进的 RetinaNet-CPN 网络, 能很好提升人体关键点整体检测精度.

1 算法流程

算法流程如图 1 所示: 对视频流切片后, 先用改进的 RetinaNet 网络对该图片进行目标检测, 接着用 Laplacian 算子对每个目标框做模糊检测, 如果大于设定的阈值, 则要用本文的方法估算出模糊核并用其恢复出清晰的图片. 最后用引入注意力机制的 CPN 网络进行关键点检测.

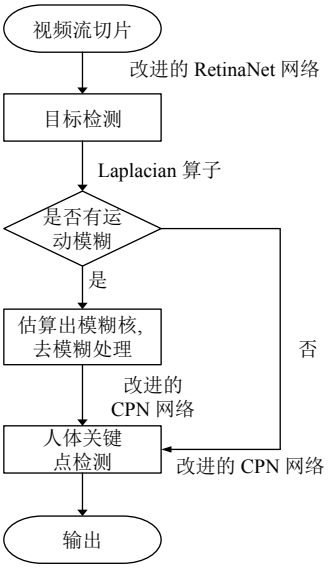

图 1 算法流程图

2 目标检测网络

综合时效性和准确率考虑, 本文选用单阶段中效果较好的 RetinaNet 做目标检测网络. 如图 2 所示, 为了充分利用各层的语义信息, RetinaNet 采用的是 ResNet+FPN[12] 结构. FPN 网络详细结构如图 3 所示, 它主要由 P_3–P_7 构成, ResNet 的 C_3–C_5 这 3 层两倍上采样后与下层经过 1×1 的卷积相加, 就成了 FPN 的 P_3–P_5 层, 采用 1×1 卷积核的目的是降低通道数, C_5 经过卷积得到 P_6, P_7 是 C_6 经过激励函数后卷积所得. 每个融合层后接一个分类网络和一个位置回归网络. 每个 FPN 特征层都有 3 种不同大小, 3 种不同长宽比例的 anchor.

单阶段目标检测效果不及双阶段的主要原因是: 负样本比例远高于正样本. 这些负样本对网络学习是没有效果的, 而双阶段利用 RPN 网络使得正负样本控制在一定比例, 学习效果更好. RetinaNet 优于其他单阶段网络主要原因是它引入平衡因子来抑制易分样本的损失权重, 损失函数 Focal Loss 的表达式如下:

$$L_{fl}=\begin{cases}-\alpha(1-y')^{\gamma}\lg y', & y=1\\ -(1-\alpha)y'^{\gamma}\lg(1-y'), & y=0\end{cases} \quad (1)$$

实验发现当 α 取 0.25, γ 取 2 时检测效果最好.

RetinaNet 用 IOU (两框的交并比) 作为预测框和真实框评价函数并用它来划分正负样本. 但 IOU 没有将两框的距离考虑进去, 当两个框没有重合即 IOU 为 0 时, 此时没有梯度回传, 不能调整模型参数; 另一方面如图 4 所示, 尽管它们的 IOU 是一样的, 但 IOU 不能很好地反映其重合的优劣性, 第一幅图效果最好, 最后一幅图最差.

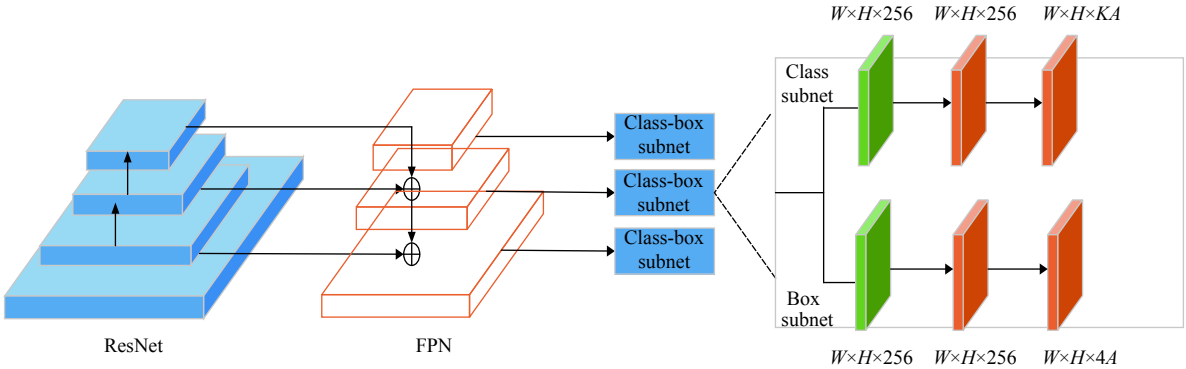

图 2　RetinaNet 网络结构

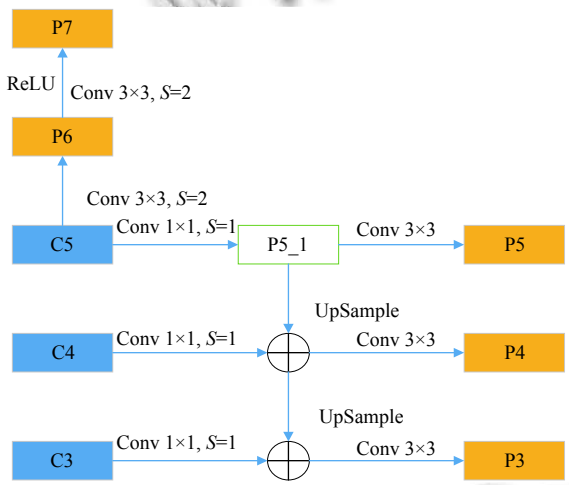

图 3　RetinaNet FPN 结构

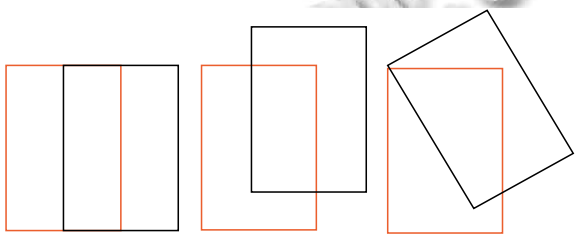

图 4　目标框重叠

DIOU[13] 同时考虑了预测框与实际框的欧式距离, 重叠率和两框的大小, 能够很好地解决上述问题, 计算公式如下:

$$DIOU=IOU-\frac{\rho^2(b,b^{gt})}{c^2} \quad (2)$$

其中, b, b^{gt} 分别代表预测框和真实框的中心点, $\rho^2(b,b^{gt})$ 表示这两者的欧氏距离的平方, c 代表的是能够同时包含预测框和真实框的最小闭包区域的对角线距离. 当两框的 IOU 为 0 时, DIOU 可以指导预测框往哪个方向调整, 能得到更佳更稳定的回归.

3　运动模糊检测及图像复原

如图 5 所示, 运动模糊主要是在摄像机曝光时刻内, 目标快速移动造成的. 对于一幅图片来说, 如果该图片中的高频分量较少则可认为它是模糊的. 拉普拉斯算子是一种常用的检测图片是否模糊的二阶微分线性算子, 公式如式 (3) 所示, 其检测过程是: 先将每个目标框的图片 resize 成固定大小的图片, 灰度化后用 Laplacian 算子滤波, 计算其方差. 由于模糊图片很难提取边缘, 因此方差较小, 如果计算出的值小于阈值则认为它是模糊照片.

$$\nabla f(x,y)=\frac{\partial^2 f}{\partial x^2}+\frac{\partial^2 f}{\partial y^2}=f(x+1,y)+f(x-1,y)+f(x,y+1)+f(x,y-1)-4f(x,y) \quad (3)$$

图 5　运动模糊图片

运动模糊图像在忽略噪声情况下可以认为由原清晰图像与一个模糊核卷积而成, 所以关键就是估算出模糊核的方向及大小, 然后用逆傅里叶变换即可得到清晰的图像. 由于摄像头的曝光时间比较短, 目标在这段时间里的运动可看成是匀速直线运动. 设曝光时长为 T, 目标沿位移方向运动了 L 像素即模糊核大小为 L, $x_0(t)$和$y_0(t)$是位移方向分解的两个运动分量, 模糊图像 $g(x, y)$ 可由清晰图像 $f(x, y)$ 积分得到, 其关系式如式 (4) 所示:

$$g(x,y)=\int_0^T f(x-x_0(t),y-y_0(t))dt \tag{4}$$

对式 (4) 进行二维傅里叶变换:

$$\begin{aligned}G(u,v)&=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} g(x,y)e^{-j2\pi(ux+vy)}dxdy\\&=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\left(\int_0^T f(x-x_0(t),y-y_0(t))dt\right)e^{-j2\pi(ux+vy)}dxdy\\&=\int_0^T\left[\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x-x_0(t),y-y_0(t))e^{-j2\pi(ux+vy)}dxdy\right]dt\end{aligned}$$

令 $F(u,v)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)e^{-j2\pi(ux+vy)}dxdy$, 积分换元后有:

$$\begin{aligned}G(u,v)&=F(u,v)\int_0^T e^{-j2\pi(ux_0(t)+vy_0(t))}dt\\&=F(u,v)H(u,v)\end{aligned} \tag{5}$$

则有:

$$H(u,v)=\int_0^T e^{-j2\pi(ux_0(t)+vy_0(t))}dt \tag{6}$$

设目标在 x 轴和 y 轴方向的位移分量分别为 a 和 b, 则$x_0(t)=\frac{a}{T}$, $y_0(t)=\frac{b}{T}$. 代入式 (6) 可得:

$$H(u,v)=\frac{T\sin(\pi(ua+vb))}{\pi(ua+vb)}e^{-j2\pi(ua+vb)} \tag{7}$$

设图像尺寸为$M\times N$, 将它表示成离散形式得:

$$|H(u,v)|=\left|\frac{T\sin\left(\pi\left(\frac{ua}{M}+\frac{vb}{N}\right)\right)}{\pi\left(\frac{ua}{M}+\frac{vb}{N}\right)}\right| \tag{8}$$

图 6 是加入 45°模糊核后的频谱图.

当$\frac{ua}{M}+\frac{vb}{N}$为非 0 整数时, $|H(u,v)|$频谱值为 0, 因此在频域图中会出现等间距的平行的条纹[14](如图 6 所示). 设频谱亮纹偏移角, 实际位移的偏移角分别为φ和θ, 有$\tan\varphi=-\frac{aM}{bN}$, $\tan\theta=\frac{b}{a}$. 它们之间的关系如式 (9) 所示:

$$\tan\varphi*\tan\theta=-\frac{M}{N} \tag{9}$$

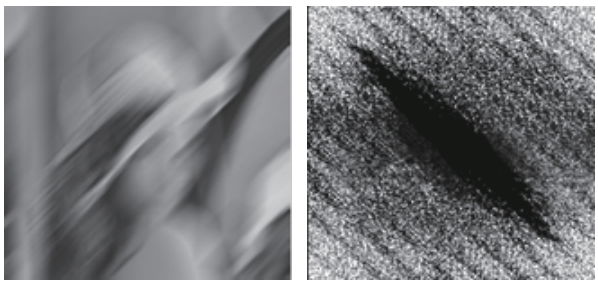

图 6 运动模糊频谱图

本文采用 Radon 变换[15] 来检测亮条纹的偏移角, 然后根据式 (9) 便可计算出模糊核的角度 α.

为了得到更加精确的模糊核大小, 将频谱绕中心点旋转 α 后, 计算出相邻暗条纹间距 d, 让$|H(u)|$的值为 0, 得到模糊核长度为 N/d.

参考文献 [16], 上述去运动模糊算法步骤总结如下:

(1) 将目标框图片灰度化后得到频谱幅度图$|G(u,v)|$.

(2) 为了使条纹表现得更加明显, 将中心点由左上角转变成频域矩形的中心点位置.

(3) 估算模糊核. 用 Radon 变换找出变换矩阵最大值的列数即为频谱条纹偏移角, 根据式 (9) 便可得到模糊核角度; 算出相邻暗纹间距, 即可得到模糊核尺度.

(4) 得到模糊核后, 经过逆傅里叶变换即可恢复出原清晰图像.

4 人体关键点检测

RetinaNet 网络检测出图片中的每个人后, 如图 7 所示, 将每个目标框裁剪后作为输入送入关键点检测网络, 人体关键点检测选用的是 CPN 网络 (Cascaded Pyramid Network), 如图 8 所示, 它主要由 GlobalNet 和 RefineNet 两个子网络组成. GlobalNet 用 ResNet 提取特征, 其中 C2–C5 这 4 个特征层采用 FPN 的结构来进行特征融合. GlobalNet 对于一些比较容易被检测到的、未被遮挡的如眼睛等关键点检测效果较好, 而对于一些难检测的, 遮挡严重的关键点检测效果不佳, 需

要依赖于 RefineNet. RefineNet 接在 GlobalNet 后面, 在 GlobalNet 每层输出后加一定个数的 bottleneck (即 residual block), 然后经过不同倍数的上采样后通过 concat 层 (其作用是将多个特征图在某个维度上进行拼接) 融合各层语义信息, 最后再经过一个 bottleneck 输出结果. 同时 GlobalNet 也采用在线难挖掘技术, 忽略易分样本的损失值, 主要计算几个难检测点的损失值, 使得网络注重难分样本的学习, 提高准确率.

为了进一步提升 CPN 网络关键点检测的准确率, 在 ResNet 后加入 CBAM 注意力机制[17] 能够提升网络性能. 如图 9 所示, CBAM 主要由 channel attention 和 spatial attention 两部分构成. channel attention 的目的是为每个通道分配合理的权重, 为了进一步提高效率及获得更丰富的信息, 特征图通过一个 bottleneck 后经过一个并行的最大池化层和平均池化层, 然后分别进入多层感知机 (MultiLayer Perceptron, MLP), 再将两者叠加后就是 channel attention. spatial attention 更加关注于位置信息, 将 channel attention 得到的加权特征图送入 spatial attention 后同样经过一个并行的 maxpool 和 avgpool 层, 得到的结果经过一个 concat 层拼接后再经过卷积层, 即可学习每个位置对预测结果的重要程度.

图 7 人体关键点检测

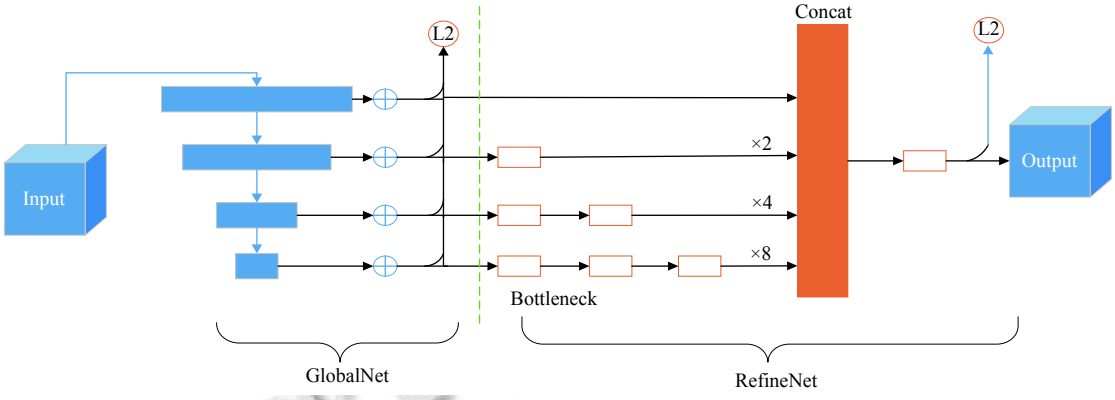

图 8 CPN 网络结构

5 实验分析

实验基于 CentOS 操作系统, Python 3.6, PyTorch 1.4, CUDA 10.1, GPU 型号为 Tesla T4, 显存为 15 GB, CPU 型号为 Intel(R) Xeon(R) Silver 4110, 2.10 GHz. 实验采用的是 COCO 数据集[18], COCO 数据集训练集有 11 万张图片, 验证集约有 5 千张图片, 测试集有 2 万多张图片, 并且该数据集对每个人的脸部和肢体共 17 个关键点做了标注. 主要做了 3 个实验: 用 DIOU 代替 IOU 做评价函数的对比实验、模糊核估算及图像复原实验和引入注意力机制 CPN[19] 的性能对比实验.

5.1 目标检测对比实验

训练时先将输入的图片 resize 成 256×256 像素, 然后用一些常用的数据增强的手法如翻转、旋转、随机裁剪等方法来提高模型的泛化能力. 实验学习率初始值为4×10^{-4}, 一共经历 15 轮训练, 评价指标采用 AP (平均准确率, Average Precision), 阈值为 0.5、0.75 的 AP. 实验测试选用的是一些常用的单阶段网络如 SSD、YOLOv3[20]、RetinaNet 等, 它们用 DIOU 代替 IOU 前后在 COCO 数据集的表现如表 1 所示.

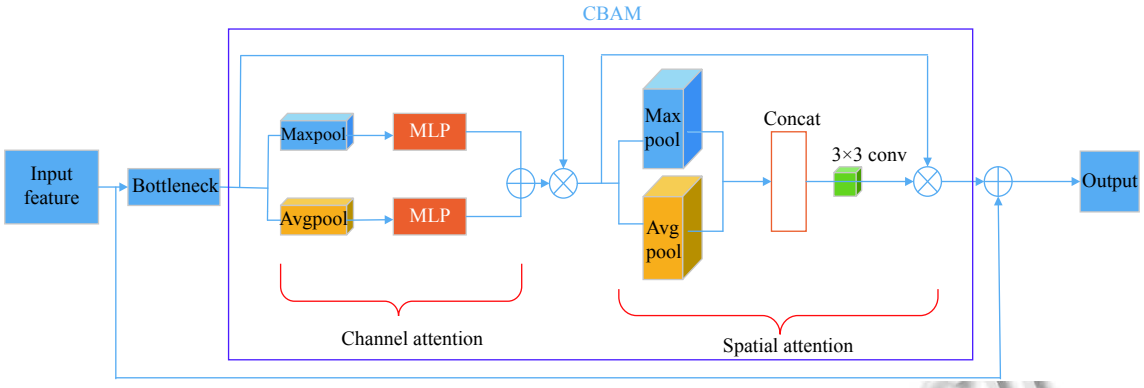

图 9 引入注意力机制的网络

表 1 用 DIOU 代替 IOU 在各单阶段网络的表现

网络	骨干网络	AP	AP_{50}	AP_{75}
SSD (IOU)	ResNet101	31.2	50.4	33.3
YOLOv3 (IOU)	Darknet	33	57.9	34.4
RetinaNet (IOU)	ResNet101+FPN	39.1	59.1	42.3
SSD (DIOU)	ResNet101	32.3	21	35.2
YOLOv3 (DIOU)	Darknet	34.1	61.2	36.7
RetinaNet (DIOU)	ResNet101+FPN	40.2	63.8	45.4

5.2 运动模糊去除实验

选取 50 张图片检测框内的人体目标加入不同的模糊核做模糊处理, 用本文的方法估算出模糊核. 图 10、图 11、图 12 是其中的 3 幅图, 其实际与估算的模糊角度, 模糊核长度如表 2 所示.

表 2 实际模糊核与估算模糊核

图片	实际模糊角度 (度)	实际模糊长度 (像素)	估算角度 (度)	估算长度 (像素)
run.jpg	30	15	31.2	15.6
farmer.jpg	45	10	44.7	8.6
woman.jpg	60	20	62.8	17.3

5.3 CPN 网络引入注意力机制前后的对比实验

关键点检测采用的是 RetinaNet+CPN 的方式, 将每个目标框 resize 成 128×256 送入 CPN 网络检测关键点. 实验学习率初始值为5×10^{-4}, 一共经历了 15 轮训练. 评价指标采用 AP、AR (平均召回率, Average Recall) 以及阈值为 0.5, 0.75 的 AP、AR, 结果如表 3 所示.

5.4 实验结果分析

将 RetinaNet 正负样本评价函数由 IOU 变成 DIOU 后, AP 由 39.1 提高到 40.2, 性能提高了 2.81%, 同时将 DIOU 用在其他单阶段网络也有不错的性能提升, 有很强的泛化能力. 50 幅图片中估算出的模糊核与实际模糊核角度、尺寸偏差均值分别为 3.7°和 2.1 pixel. 引入注意力机制后的 CPN 网络关键点检测 AP 和 AR 均有小幅提高, 分别提升 1.12% 和 0.51%.

图 10 run.jpg 去模糊前后对比图

图 11 farmer.jpg 去模糊前后对比图

图 12 woman.jpg 去模糊前后对比图

表 3 CPN 网络改进前后对比

网络	AP	AP_{50}	AP_{75}	AR	AR_{50}	AR_{75}
CPN网络	71.3	90.2	78.2	78.1	92.4	83.1
CPN改进	72.2	90.8	78.8	78.5	92.7	83.3

6 结论与展望

用本文提出的改进的 RetinaNet-CPN 网络将目标评价函数 IOU 改成 DIOU 能够有效提升目标框提取的准确率; 引入注意力机后 CPN 网络各通道和特征层分配合理的权重, 性能有一定的提升; 同时用基于 Radon 变换的估算出的模糊核与实际模糊核相差不大, 用它基本能够还原清晰图. 尽管用本文的方法使得关键点检测性能得到一定的提升, 但对遮挡严重, 人员密集的情况下的关键点检测依旧欠佳, 若再提升目标检测准确率对关键点检测效果提升有限, 接下来的研究应重点关注如何提高遮挡点预测准确性.

参考文献

1 Toshev A, Szegedy C. Deeppose: Human pose estimation via deep neural networks. 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014. 1653–1660.

2 Tompson J, Goroshin R, Jain A, *et al*. Efficient object localization using convolutional networks. Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston: IEEE, 2015. 648–656.

3 Carreira J, Agrawal P, Fragkiadaki K, *et al*. Human pose estimation with iterative error feedback. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE, 2016. 4733–4742.

4 Newell A, Yang KY, Deng J. Stacked hourglass networks for human pose estimation. Proceedings of the 2016 14th European Conference on Computer Vision (ECCV). Amsterdam: Springer, 2016. 483–499.

5 Sato K, Nagashima Y, Mano T, *et al*. Quantifying normal and parkinsonian gait features from home movies: Practical application of a deep learning-based 2D pose estimator. PLoS One, 2019, 14(11): e0223549. [doi: 10.1371/journal.pone.0223549]

6 Chen YL, Wang ZC, Peng YX, *et al*. Cascaded pyramid network for multi-person pose estimation. Proceeding of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7103–7112.

7 Sun K, Xiao B, Liu D, *et al*. Deep high-resolution representation learning for human pose estimation. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019. 5686–5696.

8 Liu W, Anguelov D, Erhan D, *et al*. SSD: Single shot MultiBox detector. Proceedings of the 14th European Conference on Computer Vision (ECCV). Amsterdam: Springer, 2016. 21–37.

9 Lin TY, Goyal P, Girshick R, *et al*. Focal loss for dense object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(2): 318–327. [doi: 10.1109/TPAMI.2018.2858826]

10 He KM, Gkioxari G, Dollár P, *et al*. Mask R-CNN. Proceedings of 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017. 2980–2988.

11 Cai ZW, Vasconcelos N. Cascade R-CNN: Delving into high quality object detection. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City: IEEE, 2018. 6154–6162.

12 Lin TY, Dollár P, Girshick R, *et al*. Feature pyramid networks for object detection. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017. 936–944.

13 Wu XW, Sahoo D, Zhang DX, *et al*. Single-shot bidirectional pyramid networks for high-quality object detection. Neurocomputing, 2020, 401: 1–9. [doi: 10.1016/j.neucom.2020.02.116]

14 孔勇奇, 卢敏, 潘志庚. 频谱预处理模糊运动方向鉴别的改进算法. 中国图象图形学报, 2013, 18(6): 637–646. [doi: 10.11834/jig.20130604]

15 Moghaddam ME, Jamzad M. Finding point spread function of motion blur using Radon transform and modeling the motion length. Proceedings of the Fourth IEEE International Symposium on Signal Processing and Information Technology, 2004. Rome: IEEE, 2004. 314–317.

16 陈健, 张欣, 陈忠仁. 基于 Radon 变换改进的运动模糊图像 PSF 参数估计算法. 软件, 2020, 41(6): 1–6. [doi: 10.3969/j.issn.1003-6970.2020.06.001]

17 Woo S, Park J, Lee JY, *et al*. CBAM: Convolutional block attention module. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 3–19.

18 Lin TY, Maire M, Belongie S, *et al*. Microsoft COCO: Common objects in context. Proceedings of the 13th European Conference on Computer Vision. Zurich: Springer, 2014. 740–755.

19 林怡雪, 高尚, 王光彩, 等. 基于改进 CPN 的人体关键点定位算法研究. 现代计算机, 2020, (12): 86–92.

20 Redmon J, Farhadi A. YOLOv3: An incremental improvement. arXiv: 1804.02767, 2018.