DOI:10.19651/j.cnki.emt.2212369

基于多任务对齐的密集行人检测算法研究*

安胜彪 李晔彤 白 宇
(河北科技大学信息科学与工程学院 石家庄 050018)

摘 要：行人检测是深度学习目标检测领域的重要分支,但密集场景中存在严重遮挡问题,给行人检测带来巨大挑战。为了缓解该问题,在 CenterNet 多任务学习模型上提出目标检测和姿态关键点检测任务对齐方法,改进后的模型为 Center_tood。首先提出分离模块:该模块将原始特征分离得到更加关注各个任务的特征;在此基础上提出任务对齐方法:通过设计对齐度量来约束损失,使模型在梯度上更大程度地向着多任务对齐的方向优化,同时利用一致性约束,使模型学习到不同任务之间的共性信息,从而对齐不同任务的特征。实验部分采用 CrowdPose 数据集训练和测试。本算法的目标检测 AP 值为 74.3%,提高了 11.5%;人体姿态关键点 AP 值为 55.8%,提高了 9.6%。实验结果验证了提出的多任务学习算法在密集场景行人检测上的有效性。

关键词：行人检测;遮挡问题;CenterNet;多任务学习;对齐损失
中图分类号：TP391 **文献标识码**：A **国家标准学科分类代码**：510.4050

Research on dense pedestrian detection algorithm based on multi-task alignment

An Shengbiao Li Yetong Bai Yu
(School of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang 050018, China)

Abstract: Pedestrian detection is an important branch of deep learning object detection field, but there are serious occlusion problems in dense scenes, which brings great challenges to pedestrian detection. To alleviate this problem, a task alignment method for target detection and attitude key point detection was proposed on the CenterNet multi-task learning model, and the improved model was Center_tood. Firstly, the separation module is proposed. This module separates the original features into the features that pay more attention to each task. On this basis, a task alignment method is proposed: the alignment measurement is designed to constrain the loss, so that the model can optimize towards the direction of multi-task alignment to a greater extent on the gradient. At the same time, the consistency constraint is used to make the model learn the common information between different tasks, so as to align the features of different tasks. In the experiment part, CrowdPose data set was used for training and testing. The AP value of the proposed algorithm is 74.3%, which increases by 11.5%. The key point AP value of human posture was 55.8%, which increased by 9.6%. Experimental results verify the effectiveness of the proposed multi-task learning algorithm in pedestrian detection in dense scenes.

Keywords: pedestrian detection;occlusion problem;CenterNet;multi-tasking learning;loss of align

0 引 言

行人检测作为计算机视觉中重要研究内容[1]有着非常广泛的应用场景,如自动驾驶、智能监控、智能机器人等,尤其在智能监控中发挥着极其重要的作用[2]。行人检测任务中常见的是密集行人检测,当遮挡现象发生时,行人目标的特征会出现大量干扰信息,基于深度学习的行人检测技术依据提取到的特征进行检测,所以遮挡现象会导致检测精度大幅度下降。并且行人目标的各个身体部位都有可能被遮挡,当行人之间发生遮挡时,特征图中单个行人目标自身的特征虽然不会有改变,但由于若干个行人目标的特征重叠在一起,特征图中的高响应的区域会被连接在一起。此时,会对检测器检测每个行人目标的边界造成很大的困难,导致误检和漏检现象发生。

收稿日期:2022-12-12
* 基金项目:国家自然科学基金(61902108)、河北省自然科学基金(F2019208305)项目资助

目前基于深度学习的行人检测算法针对密集遮挡问题，主要集中在模型结构、损失函数、非极大值抑制(non-maximum suppression，NMS)等方面进行改进，例如在模型结构的优化上，Zhou 等[3]提出 Bi-Box 方法在 Faster R-CNN[4]基础上添加了新的分支用于学习图像中人体可见部位的位置，再结合原分支增强对遮挡部位的检测能力。为了更好地解决行人与其他物体的遮挡问题，Zhang 等[5]提出的 OR-CNN 方法是基于部件检测的思路，将人体特征分为头部、左半身、右半身、大腿和小腿 5 个部位，根据几个不同部位的激活响应，对这些特征进行加权求和，提升了模型对遮挡情况的处理能力，但缺乏应对类内遮挡问题的能力。在损失函数的优化上，Wang 等[6]根据磁的同性相吸，异性相斥的原理提出了 Repulsion loss，该方法通过使用吸引力损失和排斥力损失使得预测框更靠近其所属的真实框，并远离其他的预测框，来降低其他框对检测结果的干扰，缺点是训练效果不佳，难以用于实际场景。在非极大值抑制的优化上，Liu 等[7]提出的 Adaptive NMS 算法为每个检测框预测一个密度值，模型根据密度值的大小能够自适应的调整非极大值抑制的阈值，使得该模型在密集场景中有更好的检测效果，降低了漏检率，缺点是网络训练难度大，网络很难学习到较好的密度值。

但目前基于多任务学习模型的密集行人检测研究的较少，MultiTask-CenterNet(MCN)[8]和 LSNet[9]两篇文章都为多任务模型，进行了目标检测、姿态估计和语义分割任务的训练，但都不是密集行人检测场景，并且不同任务分支的学习机制不同，学习到的特征的空间分布可能不同，单独的分支进行预测时，会导致一定程度的不对齐。因此针对上述问题，本文提出基于多任务对齐的密集场景行人检测算法研究，该研究在 CenterNet[10]模型基础上进行改进，主要贡献如下：

1)提出了一种基于层注意力机制的特征分离模块，通过动态计算目标检测与姿态关键点检测任务关注的特征来鼓励任务分解。

2)提出对齐方法，针对多任务训练不平衡问题，计算对齐度量来平衡梯度。针对多任务信息交互问题，提出一致性约束，使不同任务之间进行信息交互。

3)在 CrowdPose[11]数据集上的实验结果证明了特征分离模块和对齐方法的有效性，为多任务学习提供了一种新的任务对齐方法。此外，本研究的 Center_tood 模型目标检测任务达到了 74.3 AP，姿态关键点检测任务达到了 55.8 AP，大大超过了常用检测器，如 Mask RCNN[12]，FCOS[13]，Simple baseline[14]、Openpose[15]、MCN、LSNet、RepGT、OR-CNN。定性结果进一步验证了本文的任务校准方法的有效性。

1　CenterNet 网络模型

CenterNet 是一种端到端的基于无锚框的目标检测模型，其继承自 CornerNet[16]目标检测模型，可以很容易迁移到例如 3D 目标检测和姿态关键点检测等任务。该算法是将目标检测问题看作关键点估计问题，将图像输入到一个全卷积网络中生成热力图，该热力图的峰值为目标的中心，其他的属性，如目标尺寸、维度、3D 范围、方向和人体姿态关键点在中心点的位置则直接通过图像特征来回归(多个属性对应多个 channel)，即图像特征在每个峰值处预测目标边界框的高、宽和其他属性。

CenterNet 模型整体为编码-解码结构，由图 1 可知。CenterNet 模型由主干提取网络 ResNet-101[17]、反卷积模块进行编码，随后头网络进行解码，头网络分为两个任务，分别进行目标检测和人体姿态关键点检测，每个任务有 3 个分支，目标检测任务的 3 个分支用来计算中心点热度图、中心点的离散误差和检测框的宽高；姿态关键点任务的 3 个分支来计算骨骼点热度图、骨骼点的离散误差和骨骼点的偏移量。多任务分支意味着拥有多个损失，CenterNet 两个任务的整体损失均为以下 3 部分损失加和：

CenterNet 直接预测中心点坐标和大小，以每张图的 GroundTrue 来生成高斯热力图，公式如下，其中 $\tilde{p}_x$ 和 $\tilde{p}_y$ 为 GroundTrue 的中心点坐标：

$$Y_{xyc}=\exp\left(-\frac{(x-\tilde{p}_x)^2+(y-\tilde{p}_y)^2}{2\sigma_p^2}\right) \tag{1}$$

1) 中心点损失函数和骨骼点热度图损失函数，使用 Focal loss：

$$L_{hm}=L_{hm\_hp}=$$

$$\frac{1}{N}\sum_{xyc}\begin{cases}(1-\hat{Y}_{xyc})^{\alpha}\log(\hat{Y}_{xyc}) & if\ Y_{xyc}=1\\ (1-Y_{xyc})^{\beta}(\hat{Y}_{xyc})^{\alpha} & \text{其他}\\ \log(1-\hat{Y}_{xyc}) & \text{其他}\end{cases} \tag{2}$$

其中，N 为图像关键点个数，α 和 β 为超参数。

2) 中心点离散损失函数和骨骼点的离散误差损失函数，使用 L1 loss：

$$L_{offset}=\frac{1}{N}\sum_{p}\left|\hat{O}_{\tilde{p}}-\left(\frac{p}{R}-\tilde{p}\right)\right| \tag{3}$$

式中：$\hat{O}$ 为骨干网络输出的偏置值，p 代表目标框中心点，R 代表下采样倍数 4，$\tilde{p}=\left\lfloor\frac{p}{R}\right\rfloor$，$\frac{p}{R}-\tilde{p}$ 代表偏差值。

3) 宽和高损失函数和关键点偏移损失，使用的是 L1 loss：

$$L_{wh}=L_{hp}=\frac{1}{N}\sum_{k-1}^{N}\left|\hat{s}_{pk}-s_k\right| \tag{4}$$

各个任务整体的损失函数是以上三者之和，并分配了不同的权重：

$$L_{det} = w_{hm}L_{hm} + w_{offset}L_{offset} + w_{wh}L_{wh} \quad (5)$$

$$L_{kp} = w_{hm\_hp}L_{hm\_hp} + w_{offset}L_{offset} + w_{hp}L_{hp} \quad (6)$$

2 本文方法

本文使用 ResNet-101 为骨干网络、反卷积模块为颈部、CenterNet 模型的目标检测网络和人体姿态关键点网络作为头部网络。基于此本文做了如下改进:1)加入分离模块,使得原始特征更加关注各个任务特征;2)设计任务对齐度量来约束损失,减少关键点检测的偏差;3)利用一致性约束对齐两个任务信息交互的特征。改进后的模型如图 1 所示。

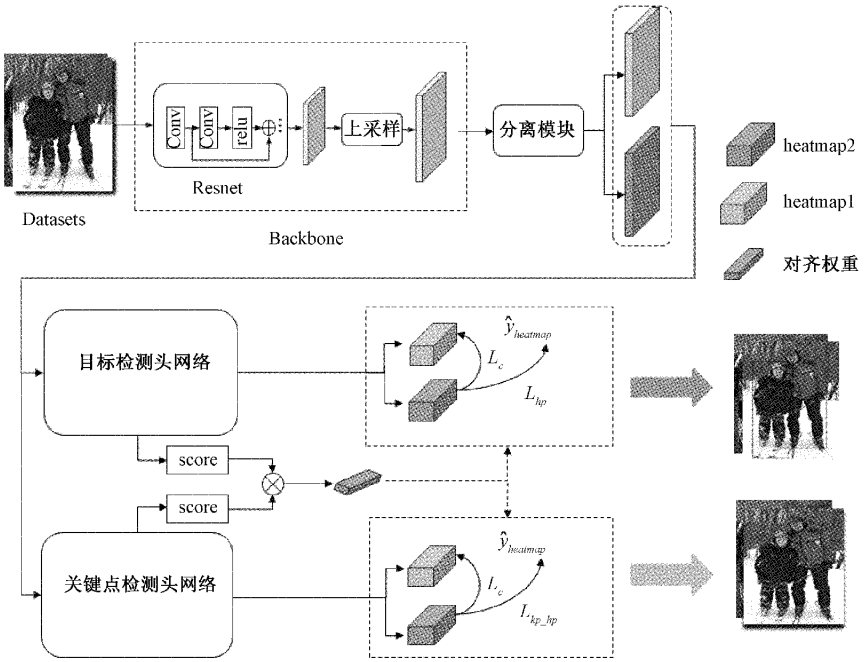


图 1 Center_tood 模型

2.1 基于 CenterNet 的分离模块

在密集场景的行人检测中,对模型的特征提取有更高的要求,但在计算各自任务时,因各个任务的关注点不同,使用提取出来的原始特征会不可避免的引入两个任务的特征冲突。于是提出层注意力机制[18],通过动态计算特定于任务的特征进行任务分解,提高密集人群检测的精度。如图 2 所示。反卷积层后首先经过全局平均池化,将输出 x^{inter} 输入到全连接层 fc_1 中,之后再使用 sigmoid 激活函数,再经过一层全连接层 fc_2,将输出的 ω 与全连接层的输出特征相乘得到最后的分离特征 x_k^{task}。

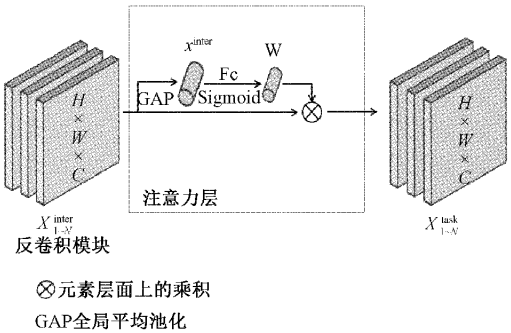


图 2 分离模块

对特定于任务的特征计算如下:

$$x_k^{task} = w_k \cdot x_k^{inter}, \forall k \in \{1,2,\cdots,N\} \quad (7)$$

其中,w_k 是第 k 个从注意力层学习到的要素,$w \in R^N$。

$$w = \sigma(fc_2(\delta(fc_1(x^{inter})))) \quad (8)$$

x^{inter} 是对 X^{inter} 使用 average pooing 后得到的。最后将输出的两个特征分别输入到任务头网络中。

2.2 基于 CenterNet 的对齐模块

针对多任务学习模型在密集场景下存在遮挡导致检测精度下降的问题,提出任务对齐方法,有效解决了因关键点被遮挡导致偏差较大出现的检测不准确问题。基于 CenterNet 模型的目标检测任务输出为 hm、wh、offset,即 hm 为中心点的位置、wh 为 bbox 的宽和高、offset 为中心点偏移量。姿态关键点检测任务输出 hm_hp、hps 和 hp_offset,即 hm_hp 为关键点的位置、hps 为关键点图像中所有关键点相对于检测框中心点的位置、hp_offset 为关键点偏移量。首先使用 Focal loss 计算 hm 预测的中心点 heatmap 和真实值 gtlabel 之间的偏差,差值代表了 bbox 预测的误差,公式如下:

$$FL(P)_{hm} = -\alpha(1-p)^{\gamma} \times y\log(p) - (1-\alpha)p^{\gamma} \times (1-y)\log(1-p) \quad (9)$$

其中,y 是样本的真实值 gtlabel,p 为模型预测某个样本为正样本的概率,对于真实标签为正样本的样本,它的概率 p 越大说明模型预测的越准确,α 和 γ 为超参数,初始值分别为 0.25 和 2。其次使用 Sigmoid 函数将计算出的

损失值压缩至(0,1)范围之间,这样就可以把 Sigmoid 看作一种"类别概率",比如 Sigmoid 的输出为 0.9 就可以解释为属于正样本的概率有 90%,公式如下:

$$S(FL(P)_{hm})_{hm}=\frac{1}{1+e^{FL(P)_{hm}}} \tag{10}$$

同理计算出 hm_hp 的误差。公式如下:

$$FL(P)_{hm\_hp}=-\alpha(1-p)^{\gamma}\times y\log(p)-(1-\alpha)p^{\gamma}\times(1-y)\log(1-p) \tag{11}$$

$$S(FL(P)_{hm\_hp})_{hm\_hp}=\frac{1}{1+e^{FL(p)_{hm\ hp}}} \tag{12}$$

最后设计偏差参数如式(13):

$$t=S_{hm}^{\alpha}\times S_{hm\_hp}^{\beta} \tag{13}$$

其中,α 和 β 被用来控制这两个任务的影响。对齐度量为 1 减去偏差参数,公式如下:

$$t_{alig}=1-t \tag{14}$$

当目标检测或姿态关键点检测的误差很大时,t_{alig} 会远大于两个任务误差都小的情况,从而迫使模型更加关注这种梯度不平衡,学习不充分的样本。同时观察到当对齐度量较小时,导致模型在反向传播时梯度过小,对此为度量设置下限 τ,当 $t_{alig}<\tau$ 时,将 τ 赋值给 t_{alig}。t_{alig} 在两个任务向任务对齐目标的联合优化中起着至关重要的作用,它鼓励网络从联合优化的角度动态关注高质量(即任务分配)锚点。对齐良好的锚预测的边界框(即具有较大的 t_{alig})通常具有精确的定位,并且这样的边界框在 NMS 期间更有可能被保留。此外,t_{alig} 选择的高质量包围框,会更仔细地加权损失,以提高训练。从高质量的包围框中学习有利于模型的性能,而低质量的包围框往往会产生大量信息较少、冗余的信号来更新模型,从而对训练产生负面影响。因此,通过关注对齐良好的锚点来提高目标检测和姿态关键点检测精度,同时减少错位锚点的影响。Focal loss 可以重新表述如下:

$$L_a=t_{alig}L_{hm\_hp}=$$

$$\frac{-t}{N}\sum_{xyc}\begin{cases}(1-\hat{Y}_{xyc})^{a}\log(\hat{Y}_{xyc}), & Y_{xyc}=1\\(1-Y_{xyc})^{\beta}(\hat{Y}_{xyc})^{\alpha}, & \text{其他}\\\log(1-\hat{Y}_{xyc}), & \text{其他}\end{cases} \tag{15}$$

式(13)两个任务差值的乘积,表示该样本在两个任务上的优劣情况,式(14)表示两任务对齐程度,使用该对齐参数来约束损失,使模型在梯度上更大程度上向着多任务对齐的方向优化。将中心点损失乘上对齐度量 t_{alig} 后目标检测任务总损失为下式:

$$L_{det}=t_{alig}w_{hm}L_{hm}+w_{offset}L_{offset}+w_{wh}L_{wh} \tag{16}$$

姿态关键点检测任务与目标检测任务损失计算同理,其乘上对齐度量之后的损失为式(16),总损失为式(17):

$$L_{kp}=t_{alig}w_{hm\_hp}L_{hm\_hp}+w_{offset}L_{offset}+w_{hp}L_{hp} \tag{17}$$

$$Loss=L_{det}+L_{kp} \tag{18}$$

2.3　一致性约束

多任务学习是一种联合学习,当多个任务并行学习时结果会相互影响。基于 Centernet 多任务学习模型的目标检测和姿态关键点为两个相关任务,相关任务中存在对于另一个任务的有用信息,故将分离模块分离出来的目标检测特征和姿态关键点特征分别输入到两个任务分支中,实现相关任务信息的传递,如图 1 所示。将分离特征分别输入到头网络中,每个头网络的热度图分支获得两个预测结果即为 hm_1、hm_2 和 hm_hp_1、hm_hp_2,之后使用均方误差(mean square error,MSE)loss 去对齐这两个输出,目的是虽然进行分离,但不同任务之间依然有共同的地方,通过输出保持一致性,使得不同任务之间共性地方得到对齐。目标检测任务使用 MSELoss 公式如下:

$$Loss_{hm}=\frac{1}{M}\sum_{0}^{m}(hm_1-hm_2)^2 \tag{19}$$

姿态关键点检测任务使用 MSELoss 公式如下:

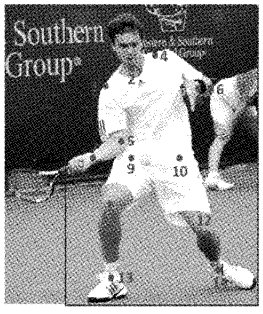
$$Loss_{hm\_hp}=\frac{1}{M}\sum_{0}^{m}(hm\_hp_1-hm\_hp_2)^2 \tag{20}$$

其中,M 为样本量。将这两个损失加入到损失总和中,为下式:

$$Loss=L_{det}+L_{kp}+t_{alig}Loss_{hm}+t_{alig}Loss_{hm\_hp} \tag{21}$$

3　实验分析

本文实验中使用了常用的行人检测数据集 CrowdPose 数据集,总共包含 20 000 张图片,80 000 行人。训练、验证、测试子集按照 5∶1∶4 的比例进行划分。CrowdPose 数据集具有目标框标注和人体关键点标注,人体关键点标注为 14 个,具体实例如图 3 所示。



1.头　2.脖子
3.右肩　4.左肩
5.右肘　6.左肘
7.右手腕　8.左手腕
9.右臀　10.左臀
11.右膝　12.左膝
13.右脚踝　14.左脚踝

图 3　人体关键点标注

3.1　评估标准

该模型有两个任务分别为目标检测和姿态关键点检测,两个任务的评估标准都为平均召回率(average recall,AR)和平均精度值(average precision,AP)。召回率(recall,R)是正样本预测正确的结果与正样本预测正确的结果和正样本预测错误的和的比值,主要反映出来的是预测结果中的漏检率。精确率(precision,P)为识别正样本图片时,正确识别样本所占的比率。召回率和精确率公式如下:

$$Precision = \frac{TP}{TP + FP} \tag{22}$$

$$Recall = \frac{TP}{TP + FN} \tag{23}$$

其中,TP 是正类判定为正类、FP 是负类判定为正类、FN 是正类判定为负类、TN 是负类判定为负类。在 Precision-Recall 曲线基础上,通过计算每个 recall 值对应的 Precision 值的平均,可以获得一个数组形式的评估为 AP 值,其公式如下:

$$AP = \sum_{i=1}^{n-1} (r_{i+1} - r_i) P_{\text{inter}}(r_i + 1) \tag{24}$$

AR 为召回 IOU 曲线下面积的两倍,公式为:

$$AR = \int_{0.5}^{1} Recall(o)\,do \tag{25}$$

AP 值和 AR 值越高说明模型检测性能越好。

3.2 实验说明

实验基于 Ubuntu18.04 系统,GeForce GTX 3090 GPU 显卡进行训练和测试,使用 CUDA11.3、pytorch1.10.1 和 Python3.8.12 的软件平台。实验使用 ResNet-101 算法官方提供的训练权重作为加入分离模块后模型的预训练权重,之后实验使用上一次训练保存好的权重,学习率设为 1.25×10^{-4},使用随机梯度下降法(stochastic gradient descent,SGD)[19]进行 50 个 epoch 训练,batchsize 设为 8,测试时,匹配置信度为 0.4。

3.3 实验结果分析

使用 CrowdPose 数据集对 Center_tood 算法与 Mask RCNN、FCOS、Simple baseline、Openpose、MCN、LSNet、RepGT、OR-CNN 进行了比较,对目标检测和姿态关键点检测的 AP 以及 AR 如表 1 所示。RepGT 和 OR-CNN 两种算法都是针对密集行人检测做的改进,效果最好,AP 值分别为 65.2%和 68.1%,AR 值为 56.3%和 56.9%;经典的目标检测算法和姿态关键点检测算法都不是针对密集行人场景,所以相比于前两个模型精度有些下降,Mask R-CNN 算法和 FCOS 算法目标检测的 AP 值为 63.7%和 64.3%,AR 值为 55.8%和 56.0%;Simple baseline 算法和 Openpose 算法姿态关键点检测 AP 值分别为 47.9%和 52.6,AR 值分别为 65.5%和 68.1%;MCN 算法、LSNet 算法和 CenterNet 算法都为多任务模型,多任务模型相比单任务模型大,精度较低,其中 MCN 算法在 CenterNet 算法上多加了语义分割任务,更多是工程优化工作,在密集行人检测中精度比 CenterNet 算法低。LSNet 算法中目标检测采取的是回归目标每个轮廓点位置的做法,CenterNet 只是回归 4 个角点,相比来说 LSNet 算法更加准确,目标检测 AP 值比 CenterNet 高 0.7%,姿态关键点检测与 CeterNet 算法相同,用一个 anchor 点和指向 17 个关键点的 17 个向量确定 pose,精度值与 CenterNet 算法相近。而 Center_tood 在同时训练多任务的同时也提高了精度,相比于目标检测精度最高的 OR-CNN 算法 AP 值提高了 6.2%,AR 值提高了 0.5%,相比于姿态关键点检测精度最高的 Openpose 算法 AP 值提高了 3.2%,AR 值提高了 2.3%。上述结果表明 Center_tood 算法不但同时并行目标检测和姿态关键点检测两个任务,并且在行人检测任务上的精度领先于目标检测和姿态关键点检测单任务基线模型与多任务基线模型。实验结果说明本文算法有效的缓解了多任务行人检测算法在密集行人检测任务的不足。

表 1 不同算法比较

方法	Bbckbone	AR@0.5		AP@0.5	
		AR_{bbox}	AR_{kp}	AP_{bbox}	AP_{kp}
Mask R-CNN	ResNet-101	0.558	—	0.637	—
FCOS	ResNet-101	0.560	—	0.643	—
Simple baseline	ResNet-101	—	0.655	—	0.479
Openpose	ResNet-101	—	0.681	—	0.526
MCN	ResNet-101	0.462	0.651	0.540	0.442
LSNet	ResNet-101	0.556	0.675	0.635	0.464
RepGT*	ResNet-101	0.563	—	0.652	—
OR-CNN*	ResNet-101	0.569	—	0.681	—
CenterNet	ResNet-101	0.555	0.671	0.628	0.462
Center_tood	ResNet-101	0.574	0.704	0.743	0.558

注:“*”为密集行人检测模型

实验 1 在 CenterNet 网络的反卷积之后加入了分离模块,当使用分离出来的 bbox 特征和 keypoints 特征分别输入到目标检测任务头和姿态关键点检测任务头时,目标检测任务 AP@0.5 值提高了 9.2%,姿态关键点任务 AP@0.5 值却下降了 11.2%,得出来分离出来的 keypoints 特征比原始特征效果差的结论。之后将原始特征输入到姿态关键点任务头中,与之前的姿态关键点特征做对比,姿态关键点任务 AP@0.5 值上升了 5.9%。从 AR 数值上升和

可视化图中可以得出，加入分离模块后，密集人群漏检情况变少，被遮挡人可以检测出来，虽然存在错检情况，但明显加入分离模块的效果更好。对比结果如表 2 所示，可视化效果图如图 4 所示。

表 2　加入分离模块对比实验

模型	分离模块			Bbox		Keypoints	
	原始特征	分离 bbox 特征	分离 keypoints 特征	AP@0.5	AR@0.5:0.95	AP@0.5	AR@0.5
Centernet	√			0.628	0.555	0.462	0.671
		√	√	0.720	0.538	0.350	0.533
	√	√		0.742	0.557	0.521	0.665

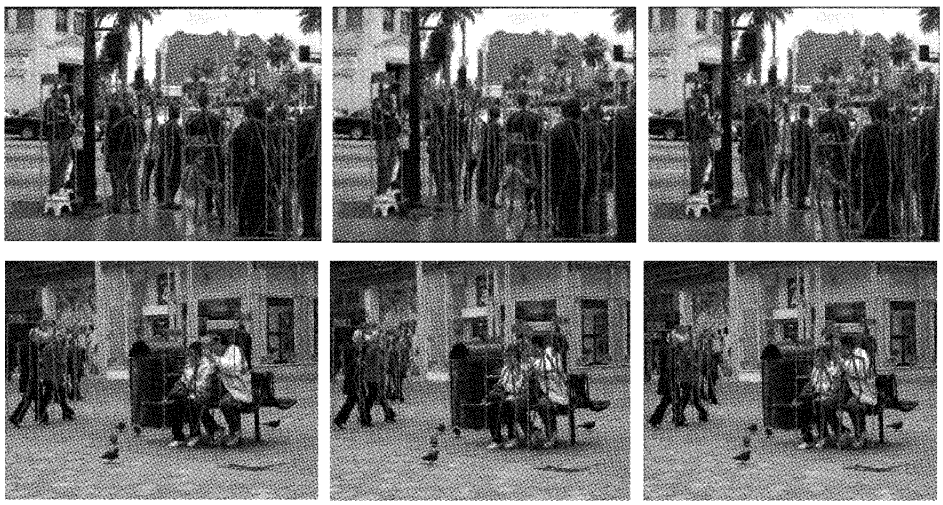
（a）原始模型　　（b）加入分离模块模型

图 4　加入分离模块效果图对比

实验 2 加入对齐度量，将式(13)中的影响因子 α 和 β 设置为 2，不做主辅任务的区分，联合训练优化两个任务。只加入对齐度量后，对齐了中心点的真实值和预测值，偏差变小了，目标检测任务 AP@0.5 值较原始模型提升了 6.9%，对齐了姿态关键点的真实值和预测值，姿态关键点检测 AP@0.5 值提升了 2.6%；在加入分离模块后加入对齐度量，使得各自任务更加关注特定特征的同时，减少了关键点的偏差。目标检测任务 AP@0.5 值较原始模型提升了 9.6%，姿态关键点任务 AP@0.5 值提升了 8.2%，比只加入对齐度量的 AP 值更高。从 AR 数值上升和可视图中可以得出，只加入对齐度量的模型关键点检测效果变好，但仍然存在漏检情况，而基于分离模块加入对齐度量后，漏检情况变少，姿态关键点检测的更加准确。对比结果如表 3 所示，可视化效果图如图 5 所示。

表 3　加入对齐损失对比实验

模型	分离模块			对齐	Bbox		Keypoints	
	原始特征	分离 bbox 特征	分离 keypoints 特征	度量	AP@0.5	AR@0.5:0.95	AP@0.5	AR@0.5
Centernet	√				0.628	0.555	0.462	0.671
	√			√	0.697	0.571	0.488	0.673
		√	√	√	0.729	0.572	0.403	0.599
	√		√	√	0.724	0.571	0.544	0.697



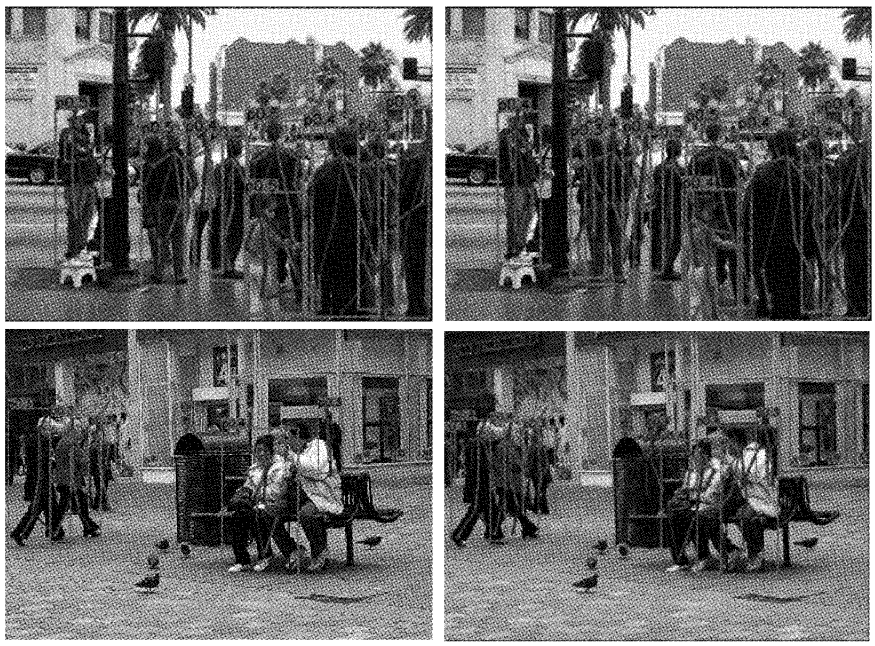
(a) 原始模型 (b) 只加入对齐loss (c) 在分离模块上加入对齐loss

图 5 加入对比 loss 效果图对比

实验 3 将分离模块分离出来的 bbox 特征和 keypoints 特征分别输入到两个任务检测头中,进行信息交互,将两个任务的特征进行了补全,再对各个检测头的两个输出进行一致性约束,计算了两个特征的偏差,使得模型向着更小偏差方向优化。实验结果表明目标检测任务 AP@0.5 值较原始模型提升了 11.5%,姿态关键点检测 AP@0.5 值提升了 9.6%。从 AR 数值上升和可视化图中可以得出,加入一致性损失后,模型漏检情况变好,姿态关键点检测也较为准确。对比结果如表 4 所示,可视化效果图如图 6 所示。

表 4 一致性约束对比实验

模型	分离特征			一致性约束	Bbox		Keypoints	
	原始特征	分离 bbox 特征	分离 keypoints 特征		AP@0.5	AR@0.5:0.95	AP@0.5	AR@0.5
Centernet	√				0.628	0.555	0.462	0.671
		√	√	√	0.715	0.569	0.416	0.619
	√	√		√	0.743	0.574	0.558	0.704



(a) 原始模型 (b) 加入一致性loss

图 6 加入一致性约束效果图对比

4　结　　论

本文提出了一种面向拥挤人群场景的密集行人检测算法，在多任务学习模型 CenterNet 中添加了分离模块、对齐度量，进行了信息传递和一致性约束。分离模块是将原始特征中特定于任务的特征分离出来，使特征更加关注于各个任务特征，模型学习效果更好；对齐度量缓解了真实值与预测值之间的偏差，将关键点进行了对齐，并且使模型更加关注梯度不平衡；多任务学习各个任务有较强相关性，学习到的特征对另一个任务有用，进行任务之间的信息传递，补全各任务的信息，并加入一致性约束，将各个任务的两个输出进行对齐，完成了信息交互。该算法使用非常具有挑战性的密集人群数据集 CrowdPose 进行训练和测试，最后的实验结果验证了本文方法在密集场景下的有效性。未来将会针对行人检测的小目标问题和部署问题进一步改进，考虑优化骨干网络和特征融合网络，在提高算法 AP 的基础上适当减少模型大小和训练时长，更好的适应实际工程中的应用。

参考文献

[1] 刘毅，于畅洋，李国燕，等. UAST-RCNN：遮挡行人的目标检测算法[J]. 电子测量与仪器学报，2022，36(12)：168-175.

[2] 王颖，金若辰，金志刚. 支持行人检测的智能车载监控终端[J]. 电子测量技术，2019，42(6)：17-21，DOI：10.19651/j.cnki.emt.1802234.

[3] ZHOU C，YUAN J. Bi-box regression for pedestrian detection and occlusion estimation[C]. Proceedings of the European Conference on Computer Vision(ECCV)，2018：135-151.

[4] REN S，HE K，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.

[5] ZHANG S，WEN L，BIAN X，et al. Occlusion-aware R-CNN：Detecting pedestrians in a crowd[C]. Proceedings of the European Conference on Computer Vision(ECCV)，2018：637-653.

[6] WANG X，XIAO T，JIANG Y，et al. Repulsion loss：Detecting pedestrians in a crowd[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition，2018：7774-7783.

[7] LIU S，HUANG D，WANG Y. Adaptive nms：Refining pedestrian detection in a crowd[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition，2019：6459-6468.

[8] HEUER F，MANTOWSKY S，BUKHARI S，et al. Multitask-centernet (mcn)：Efficient and diverse multitask learning using an anchor free approach[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision，2021：997-1005.

[9] DUAN K，XIE L，QI H，et al. Location-sensitive visual recognition with cross-iou loss[J]. ArXiv Preprint，2021，ArXiv：2104.04899.

[10] ZHOU X，WANG D，KRAHENBUHL P. Objects as points[J]. ArXiv Preprint，2019，ArXiv：1904.07850.

[11] LI J，WANG C，ZHU H，et al. Crowdpose：Efficient crowded scenes pose estimation and a new benchmark[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition，2019：10863-10872.

[12] HE K，GKIOXARI G，DOLLAR P，et al. Mask r-cnn[C]. Proceedings of the IEEE International Conference on Computer Vision，2017：2961-2969.

[13] TIAN Z，SHEN C，CHEN H，et al. Fcos：Fully convolutional one-stage object detection[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision，2019：9627-9636.

[14] XIAO B，WU H，WEI Y. Simple baselines for human pose estimation and tracking[C]. Proceedings of the European Conference on Computer Vision(ECCV)，2018：466-481.

[15] 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，2017：7291-7299.

[16] LAW H，DENG J. Cornernet：Detecting objects as paired keypoints[C]. Proceedings of the European Conference on Computer Vision(ECCV)，2018：734-750.

[17] HE K，ZHANG X，REN S，et al. Deep residual learning for image recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition，2016：770-778.

[18] VASWANI A，SHAZEER N，PARMAR N，et al. Attention is all you need[J]. Advances in Neural Information Processing Systems，2017，30.

[19] RUDER S. An overview of gradient descent optimization algorithms[J]. ArXiv Preprint，2016，ArXiv：1609.04747.

作者简介

安胜彪，硕士，副教授，主要研究方向为集成电子系统和集成电路的研究。

E-mail：33588253@qq.com

李晔彤，硕士研究生，主要研究方向为计算机视觉。

E-mail：1207376834@qq.com

白宇，博士，讲师，主要研究方向为信息物理系统(CPS)、同步系统、深度学习、基于模型的系统设计、形式化方法。

E-mail：baiyu@hebust.edu.cn