Anchor free与Anchor base算法结合的拥挤行人检测方法

谢明鸿[①] 康 斌[①] 李华锋[①②] 张亚飞[*①②]

[①](昆明理工大学信息工程与自动化学院 昆明 650504)

[②](昆明理工大学云南省人工智能重点实验室 昆明 650504)

摘 要：由于精度相对较高，Anchor base算法目前已成为拥挤场景下行人检测的研究热点。但是，该算法需要手工设计锚框，限制了其通用性。同时，单一的非极大值抑制(NMS)筛选阈值作用于不同密度的人群区域会导致一定程度的漏检和误检。为此，该文提出一种Anchor free与Anchor base检测器相结合的双头检测算法。具体地，先利用Anchor free检测器对图像进行粗检测，将粗检测结果进行自动聚类生成锚框后反馈给区域建议网络(RPN)模块，以代替RPN阶段手工设计锚框的步骤。同时，通过对粗检测结果信息的统计可得到不同区域人群的密度信息。该文设计一个行人头部-全身互监督检测框架，利用头部检测结果与全身的检测结果互相监督，从而有效减少被抑制与漏检的目标实例。提出一种新的NMS算法，该方法可以自适应地为不同密度的人群区域选择合适的筛选阈值，从而最大限度地减少NMS处理引起的误检。所提出的检测器在CrowdHuman数据集和CityPersons数据集进行了实验验证，取得了与目前最先进的行人检测方法相当的性能。

关键词：行人检测；Anchor base；Anchor free；非极大值抑制

中图分类号：TN911.73; TP391.41 **文献标识码**：A **文章编号**：1009-5896(2023)05-1833-09

DOI: 10.11999/JEIT220444

Crowded Pedestrian Detection Method Combining Anchor Free and Anchor Base Algorithm

XIE Minghong[①] KANG Bin[①] LI Huafeng[①②] ZHANG Yafei[①②]

[①](*Faculty of Information Engineering and Automation*, *Kunming University of Science and Technology*, *Kunming* 650504, *China*)

[②](*Key Laboratory of Artificial Intelligence of Yunnan Province*, *Kunming University of Science and Technology*, *Kunming* 650504, *China*)

Abstract: Due to its relatively higher accuracy, the Anchor base algorithm has become a research hotspot for pedestrian detection in crowded scenes. However, the algorithm needs to design manually anchor boxes, which limits its generality. At the same time, a single Non-Maximum Suppression (NMS) screening threshold acting on crowd areas with different densities will lead to a certain degree of missed detection or false detection. To this end, a dual-head detection algorithm combining Anchor free and Anchor base detectors is proposed. Specifically, the Anchor free detector is used to perform rough detection on the image, and the coarse detection results are automatically clustered to generate anchor frames and then fed back to the Region Proposal Network (RPN) module, instead of manually designing the anchor frames in the RPN stage. Meanwhile, the density information of the population in different regions can be obtained through the statistics of the rough detection result information. A pedestrian head-whole body mutual supervision detection framework is designed, and the head detection results and the whole body detection results supervise each other, so as to reduce effectively the suppressed and missed target instances. A novel NMS method is proposed, which can adaptively select appropriate screening thresholds for crowd regions of different densities, thereby minimizing false detections caused by NMS process. The proposed detector is experimentally validated on the CrowdHuman dataset and the CityPersons dataset, achieving comparable performance to current state-of-the-art pedestrian detection methods.

Key words: Pedestrian detection; Anchor base; Anchor free; Non-Maximum Suppression (NMS)

收稿日期：2022-04-14；改回日期：2022-08-31；网络出版：2022-09-08

*通信作者： 张亚飞 zyfeimail@163.com

1 引言

行人检测是计算机视觉领域一项非常重要的工作，它为行人重识别[1]、行人跟踪[2]以及步态识别[3]等多个视觉研究领域提供了重要基础和技术支撑，并在自动驾驶、智慧城市等领域得到了广泛的应用。现有的行人检测器[4]在非拥挤场景中具有良好的表现，但是，在商业街、机场等人群拥挤的场景中，行人目标检测仍然面临极大挑战。

在拥挤场景中，行人检测主要面临两个挑战：(1)其他目标和背景信息对行人目标的遮挡会导致检测器性能急剧下降。针对这一困难，Zhou等人[5]首次使用行人可见信息和完整信息来挖掘更好的行人特征信息。Wang等人[6]通过设计一种新的损失函数来解决遮挡问题。Chi等人[7]与陈勇等人[8]利用人头信息作为解决问题的线索。虽然这些利用行人部分信息的检测方法取得了一定效果，但是都没有很好地挖掘到行人部分与全身之间的对应关系。(2)NMS处理时，采用单一阈值设定会导致大量误检和漏检。因为拥挤场景中不同区域的人群密度差异很大，阈值设定较低会漏检高度重叠的目标实例，而阈值设定过高会带来更多的假阳例。针对这一问题，Liu等人[9]提出一种采用动态阈值的非极大值 抑制(Non-Maximum Suppression, NMS)算法，该算法通过一个密度子网预测每个位置的密度，根据所得到的不同密度值为NMS处理设置适合的交并比(Intersection over Union, IoU)阈值。Huang等人[10]尝试利用行人目标之间的密度差异信息寻找适合不同密度人群的IoU阈值，以减少NMS处理导致的检测失败。但是，密度估计是一项非常复杂和困难的任务，而将不同密度人群区域精确匹配到对应的IoU则更加困难。此外，目前大部分针对行人检测的工作都是基于Anchor base算法[11]框架实现的，可以得到较高的检测精度。但这类算法往往需要根据先验信息手工设计固定大小的锚框，极大限制了检测器的鲁棒性和通用性。

针对上述问题，本文提出了一种Anchor free与Anchor base检测器相结合的双头行人检测算法。首先，利用Anchor free检测器对检测对象先进行一次粗检测，对得到的预测框进行K-means聚类，为RPN模块[11]选取合适的锚框大小提供预设信息，从而代替了RPN模块中原本需要手工设计锚框的步骤。同时，根据粗检测结果，将人群区域划分为多个不同密集程度的区域。为了降低漏检率，设计了一种行人头部-全身互监督检测框架，通过构建头部-全身锚框对的强绑定关系，有效弥补了全身检测与头部检测各自的缺陷。此外，提出了一种新的NMS算法，称为Stripping-NMS。通过预先剥离Anchor free检测头[12]粗检测所得到的高置信度预测框，从而避免了与该固定框重叠程度较高的预测框之间互相抑制造成的漏检。同时，针对划分好的人群密度区域自适应地采用合适的IoU阈值进行NMS处理。

2 网络模型设计

本文模型建立在ResNet-50网络的基础上，其总体结构如图1所示。模型主要包括3个部分：(1)RPN与Anchor free检测头交叉网络。该网络主要为检测阶段生成高质量的候选建议框并判断人群密度差异。(2)互监督检测模块。该模块可以充分利用头部部分特征来辅助完成全身框的检测。(3)用于后处理的Stripping-NMS算法。

2.1 RPN与Anchor free检测头交叉网络

本文方法的网络框架由两个阶段组成，但与传统的两阶段算法不同的是，本文将Anchor free检测头与RPN模块结合起来作为Anchor base算法的第1阶段。首先，利用Anchor free检测头对目标对象进行一次粗检测。为了提高效率，Anchor free检测头作用于网络中较为深层的特征图。然后，统计所有预测框宽高比，通过K-means聚类算法对这些预测框进行自动聚类，聚类结果反馈给RPN模块作为锚框预设信息的依据，从而取代了RPN阶段中原本需要手工设计锚框的步骤。Anchor free检测器损失函数使用与CenterNet相同的损失函数[12]。整个第1阶段RPN模块的损失函数最终设计为

$$L_{\mathrm{RPN}} = L_{\mathrm{cls}} + L_{\mathrm{reg}}^{\mathrm{f}} + L_{\mathrm{reg}}^{\mathrm{h}} \tag{1}$$

其中，L_{cls}为前景和背景分类的Focal loss[13]，$L_{\mathrm{reg}}^{\mathrm{f}}$和$L_{\mathrm{reg}}^{\mathrm{h}}$分别是身体边界框和头部边界框的回归损失，采用Smooth L1损失。

在以往的工作中，对人群密度估计困难的主要原因在于在输出最终的预测结果之前对于不同区域中目标实例的数目完全未知。因此，本文利用Anchor free 检测头的结果来估计不同区域的人群密度。具体地，首先将给定图像等分为8个区域，然后统计了每个区域内的粗检测结果中横纵方向上相邻行人目标之间的IoU，无遮挡目标框不计入统计范围内，然后计算每个区域内所有行人之间的平均IoU作为该区域内行人的平均遮挡度。然后依据每个区域的平均遮挡度将整个区域划分为5类区域(稀疏及正常A_0、1级拥挤A_1、2级拥挤A_2、3级拥挤A_3、极致拥挤A_4)。在实验过程中我们发现，如果目标恰好位于两块区域分割处，划分时可能由于目标被切割而损害其特征信息的完整性。因此，本

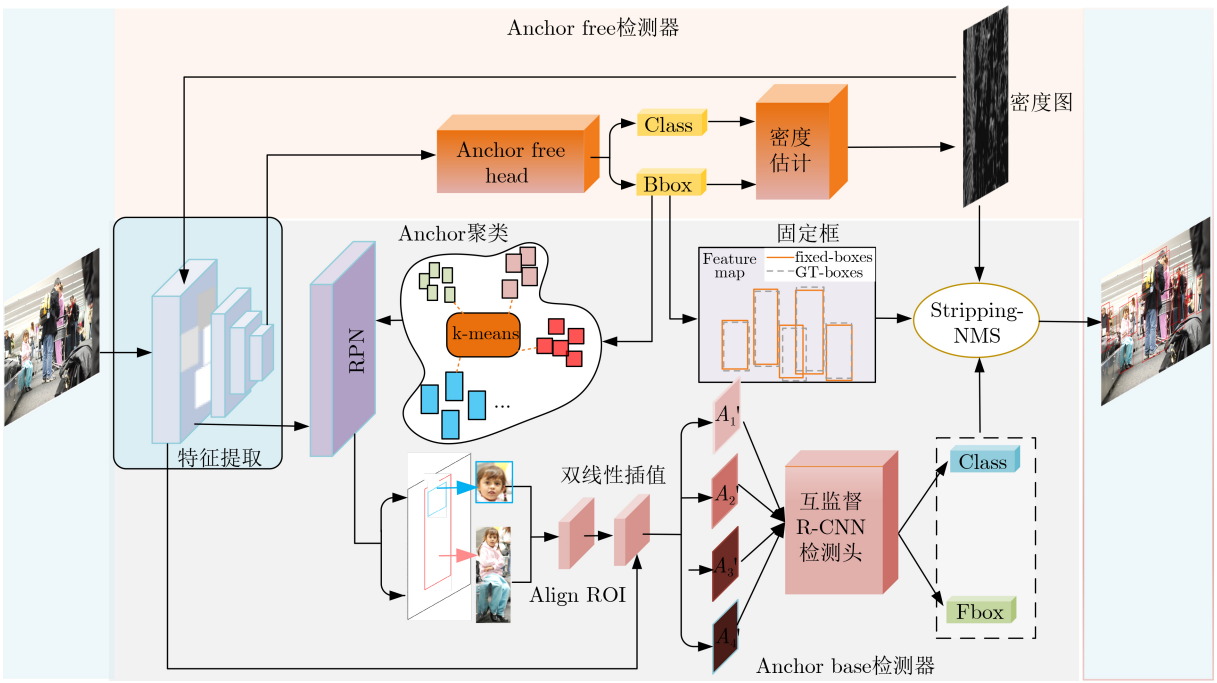


图 1 本文模型的总体结构

文使用每个区域中最左端目标的左上角横坐标及最右边目标右下角的横坐标作为分割边界。实验发现，对于稀疏区域和正常区域中的目标实例设置NMS阈值为标准的0.5具有最佳表现[9]，Anchor free检测头完全可以检测出该区域中所有目标实例。所以对于这两类区域不再送入后续检测器中处理，这将在一定程度上减少后续检测器的计算开销。

2.2 互监督检测框架

经过第1阶段处理后可得到给定图像不同区域的密度图。通过实验发现，在高密集场景下，检测器对于靠近前景的目标检测能力更强，对于景深程度较深的目标则较弱。因此，本文在将特征信息送入互监督R-CNN(Region Convolutional Neural Networks)检测头之前，对拥挤区域的特征信息进行了双线性插值以提高其表征能力，帮助检测器更好地区分高遮挡目标实例的边界。为了进一步提高检测器对困难样本的检测能力，本文设计了一个行人头部-全身互监督检测框架，如图2。为了更精准地预测行人头部，同时也为了提高网络的整体效率，本文依据统计得到的全身锚框与头部锚框的位置及比例关系(头部框与全身框的平均宽度比为2:5，平均高度比为1:5)，选取全身锚框的固定位置区域作为头部锚框，并与全身锚框建立成对关系组成锚框对。然后，利用该锚框对同时对每个行人的头部偏移和全身偏移进行回归。

可以发现，头部检测对于小目标行人检测能力更弱而对遮挡行人目标检测能力更强；而全身检测对于遮挡行人目标检测能力更弱，对小目标行人检测能力更强。因此，得到的行人头部建议框和全身建议框中有部分头部框和全身框无法保持成对关系。注意，此时的结果尚未经过NMS处理。为了防止NMS对高重叠实例预测结果的抑制，首先进行第1次头部框与全身框的匹配。对第1次匹配的结果进行NMS处理，并将NMS处理得到预测结果进行第2次匹配。将第2次匹配后头部框和全身框依旧保持成对关系的预测结果，直接返回该行人的全身框作为最终预测结果。两次匹配完成会有一些没有和全身框匹配的头部预测框。这可能是因为它对应的全身建议框在NMS处理时被抑制掉了，也可能它本身就是假阳例。为了减少被抑制与漏检的目标实例，依据两次匹配结果，对于第1次匹配时有对应的全身预测框而第2次匹配时没有对应全身预测框的头部预测结果，直接利用头部预测框把抑制掉的全身框进行召回，对于两次匹配都没有对应全身预测结果的头部预测框则直接去除。

本文在CIoU(Complete-IoU)损失函数[14]的基础上引入了人群区域密度信息，设计了一种新的更适合密集场景的损失函数，称为CD-IoU(Complete and Density-IoU)损失，作为第2阶段的回归损失函数。CD-IoU损失函数为

$$L_{\mathrm{CD-IoU}} = 1 - \mathrm{IoU} + \frac{\rho^2(b, b^{\mathrm{gt}})}{c^2} + \alpha v + \beta d \quad (2)$$

其中，IoU代表预测框与标签的交并比，b和b^{gt}分别表示预测框和真实标签的中心点，$\rho(\cdot)$表示欧氏距

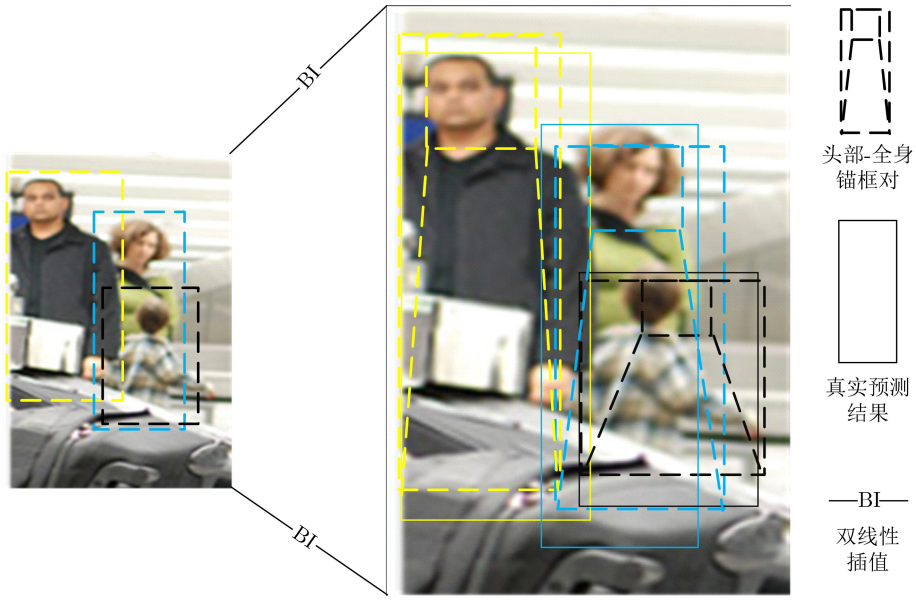


图 2 行人头部-全身互监督检测框架

离，c是覆盖两个边界框的最小封闭框的对角线长度，并通过引入额外的超参数α和v反映预测框宽高比之间的差异。d表示目标所在区域的密度，β是控制该目标位于哪个密集区域的超参数。

$$\nu = \frac{\pi}{4}\left(\arctan\frac{w^{\mathrm{gt}}}{h^{\mathrm{gt}}} - \arctan\frac{w}{h}\right)^2 \tag{3}$$

其中，w^{gt}和w分别代表真实标签与预测框的宽度，h^{gt}和h分别代表真实标签与预测框的高度。

$$\alpha = \frac{v}{(1-\mathrm{IOU})+v} \tag{4}$$

$$d = \frac{\sum_{i=1}^{n}\rho(x_i, x_{i-1})}{n} \tag{5}$$

其中，n代表该区域中总的目标数目，x为预测结果的中心点横坐标。

2.3 Stripping-NMS算法

基于Anchor free检测头的结果与划分好的密度区域，本文提出一种新的NMS算法，称为 String NMS。具体地，通过密度估计将整张图像划分为5类不同的区域。对于稀疏和正常密度的人群区域，选择传统的IoU=0.5作为NMS筛选阈值[7]。对于其他4类不同密度的区域$[A_1, A_2, A_3, A_4]$，依据密集程度选择适合的NMS筛选阈值。为了进一步提高模型的整体效率，本文保留所有Anchor free头得到的检测结果中置信度高于0.5的框作为固定框。此时面临一个关键的问题：须保证R-CNN检测头对于固定框对应的目标产生的预测框不会抑制与其重叠度较高的目标实例。为此，本文采取了一种新的NMS抑制策略。对于R-CNN检测头产生的一系列预测框，首先与Anchor free头产生的固定框之间进行IoU计算，此时可以选取一个相对更高的NMS阈值(0.9)来先去除掉一部分预测框，其余预测框将不再对Anchor free检测头得到的这些结果框做IoU计算。因此，可以避免与这些目标高度重叠和遮挡的目标实例的预测结果相互抑制，导致检测缺失。然后，对其余预测框采用Soft-NMS[15]抑制策略为其他目标选取合适的最终预测框。

为了更清楚地表述Stripping-NMS算法，本文将其总结为算法1。

3 实验分析

本节将介绍本文实验的具体设置、实现细节、实验所使用的数据集和相关评价指标，并对实验涉及的相关参数进行了验证分析。此外，为了评估所提出行人检测器的有效性，在CrowdHuman验证集[16]上对所提出行人探测器的每个组件进行消融分析，并就本文所提出的方法与主流的行人检测算法在CityPersons数据集[17]和CrowdHuman数据集上进行广泛的比较。

3.1 数据集和评估度量

CityPersons数据集是Cityscape的一个子集，它只包含行人注释。其中有2 975张图片用于训练，500张用于验证，另外1 575张图片测试。每幅图像中平均行人数量为7人，提供了可见区域和全身标注。CrowdHuman数据集是旷视发布的用于行人检测的数据集，图片大多来自Google搜索，旨在评估拥挤场景下的检测器。CrowdHuman数据集包括15 000张训练图片、5 000张测试图片和4 370张验证图片，如表1。在训练集和验证集中有47万个实例，平均

算法1 Stripping-NMS算法

输入:
预测得分:$S=\{s_1, s_2,\cdots, s_n\}$，全身预测框:$B_f=\{b_{f1}, b_{f2},\cdots, b_{fm}\}$，头部预测框：$B_h=\{b_{h1}, b_{h2},\cdots, b_{hn}\}$，
固定框：$B_a=\{b_{a1}, b_{a2},\cdots, b_{ai}\}$，不同的密度区域：$A=\{A_0, A_1, A_2, A_3, A_4\}$，NMS threshold: $N_{\mathrm{D}}=[0.5;0.6;0.65;0.7;0.8]$，
$B=\{b_1, b_2,\cdots, b_i\}$，$B_{\mathrm{m}}$表示最大得分框，$M$表示最大得分预测框集合，$R$表示最终预测框集合。

begin:

$R\leftarrow B_a$
while $B\neq$ empty do
　$m\leftarrow$ argmax S ;
　$M\leftarrow b_{\mathrm{m}}$
　$R\leftarrow F\cup M$; $B\leftarrow B\text{-}M$
　for b_i in B do
　　if IoU(b_{ai}, b_i) ≥ 0.9 then
　　$B\leftarrow B\text{-}b_i$; $S=S\text{-}s_i$
　　end
　end
　for A_i, $N_{\mathrm{D}i}$ in A, N_{D} do
　　if $A_i\supset b_i$ then
　　if IoU(b_m, $b_{fi/hi}$) $\geq N_{\mathrm{D}i}$ then
　　$B\leftarrow B\text{–}b_{fi}(b_{hi})$; $s_{\mathrm{i}}\leftarrow s_i\,(1\text{–IoU}(M,b_i))$
　　else:
　　$B\leftarrow B\text{–}b_{fi}(b_{hi})$; $S\leftarrow s-s_i$
　　end
　end
end
return R, S

表 1　CityPersons训练集与CrowdHuman训练集

	图像数目	人数	每张图人数	有效行人
CityPersons	2975	19238	6.47	19238
CrowdHuman	15000	339565	22.64	339565

每幅图像有20多人，并且目标实例之间存在多种不同情况的遮挡。每个人体实例都注释有头部、人体可见区域和人体边框。

本文采用对数平均误检率(log-average Miss Rate, MR)，记为MR^{-2}、平均准确率(Average Precision, AP)与召回率(Recall)作为评估本文模型性能的主要指标。MR^{-2}越小表示性能越好。AP是目标检测使用最广泛的评价指标，同时反映了检测结果的精度和召回率。AP越大，表示性能越好。Recall代表正样本被预测正确的占比，Recall值越高越好。

3.2　实验设置

本文基于Pytorch框架实现了所提出的方法。采用随机梯度下降(Stochastic Gradient Descent, SGD)优化器来优化网络。在调整输入大小的同时，保留了输入图像的原始长宽比。对于CityPersons数据集，本文在一块GTX2080Ti的GPU上优化网络，每个batch包含8张图像，初始学习率为0.02，总共训练了100 epochs，并在第50 epochs将学习率降低了1/10。对于CrowdHuman数据集，在一张GTX3090的GPU上进行训练，每个Batch包含32张图像，学习率设置为1×10^{-4}，总共训练了160个epochs，并在第80 epochs和第120 epochs将学习率降低为原来的1/10。

3.3　参数分析

3.3.1　NMS困境验证

为了验证NMS处理对行人检测的影响，本文从CrowdHuman验证集中随机选取500张图，共包含10 281个行人目标，假设检测器可以生成一个得分为1.0的精确预测框。但是，在使用IoU阈值0.5执行NMS之后，仅保留了9 332个精确的预测框。在检测中遗漏了近10%的真实实例。这表明在NMS使用相对较低的IoU阈值后，即使是性能最好的行人检测器也无法检测到所有实例。相反，在NMS中设置较高的IoU阈值可保留更多的真实阳性结果，同时也会增加假阳性实例。同样，验证集中，假设所有真实实例均具有准确的预测框，则当将NMS的IoU阈值设置为0.7时，丢失率将减少到2%。但是，较高的IoU阈值在实践中不可避免地会带来更多的误报。例如，实验验证集中，经过良好训练的基于ResNet-50的Faster R-CNN会在IoU阈值为0.7的NMS之后产生大约14 800个得分超过0.5个检测盒。但是真实实例数为10 281个，因此约4 600个预测的框是冗余或误报。

此外，在NMS处理过程中，为了给本文得到的不同密集区域选取最佳的筛选阈值，本文统计了两个数据集中行人之间的遮挡程度，如表2所示。本文将该遮挡程度作为NMS过程中选择合适IoU阈值的依据。CrowdHuman数据集中针对不同密度区域分别选择[0.5,0.6,0.65,0.7,0.8]作为不同密度区域的NMS阈值；在CityPersons数据集中使用[0.5, 0.55,0.6,0.65,0.7]作为不同密度区域的NMS阈值。

3.3.2　Anchor选择分析

为了选择合适的锚框，首先单独使用Anchor

表 2　CityPersons数据集与CrowdHuman数据集的行人遮挡程度[16]

	IoU>0.5	IoU>0.6	IoU>0.7	IoU>0.8	IoU>0.9
CityPersons	0.32	0.17	0.08	0.02	0.00
CrowdHuman	2.40	1.01	0.33	0.07	0.01

free检测器对CrowdHuman验证集进行测试，测试结果为：precision = 0.805，Recall = 0.697。然后，通过对所有检测结果进行自动聚类，得到的全身锚框比例为{(1:1),(1.5:1),(2:1),(2.5:1),(3:1)}，与其他通过手工对数据集聚类的方法[11]基本具有相同的全身锚框尺度设置。可见虽然Anchor free检测器对于CrowdHuman数据集检测精度较低，但不会因为锚框的设置不准确而影响检测器的性能。而对于头部检测结果，自动聚类的方法与手工设置的方法是完全一致的，该比例被设置为{(1:2),(1:1),(2:1)}。

3.4 消融实验

本文首先在CityPersons数据集上对所提出的方法进行了消融实验分析。为此，本文评估了所提检测器的每个组成部分，包括RPN与Anchor free检测头交叉网络、互监督检测器以及Stripping-NMS算法。受文献[18]启发，本文针对行人检测任务将改进的Faster R-CNN与特征金字塔网络(Feature Pyramid Networks, FPN) [19]结合重新实现了一个更强大的baseline。baseline在CityPersons数据集的Reasonable子集和Heavy子集上分别取得了11.74%与45.24%的MR^{-2}性能。表3为在CityPersons验证集上的消融实验结果，可以看出，本文所提出的3个部件都能持续提高行人检测器的性能。通过不断改进本文提出的方法，分别在两个子集上获得了1.26%与4.43%的MR^{-2}增益。

为了验证本文方法在拥挤场景下的性能，在CrowdHuman数据集上对所提出的方法进行了进一步消融分析。baseline方法在CrowdHuman数据集上达到了85.88%的AP，80.74%的召回率和42.73%的MR^{-2}性能。表4为CrowdHuman验证集上的消融实验结果，可以看出，在CrowdHuman数据集上本文提出的所有组件仍然能够提高行人检测器的性能，并获得了1.69%的MR^{-2}、5.37%的AP以及3.52%的Recall增益。消融实验验证了本文所提方法的有效性，并且在拥挤场景下，行人检测的误检率和召回率得到明显改善。

3.5 对比实验与分析

为了验证本文所提方法的有效性，本文首先在CityPersons数据集上与主流行人检测方法进行了对比实验，表5列出了对比方法和本文方法的结果。为了公平地进行比较，本文方法与对比方法尽可能采用了相同的基线。在以VGG-16为主干网络和输入图像大小为×1的方法中，本文的方法比其他对比实验方法都要好。在采用ResNet-50作为主干网络和使用×1.3比例尺度输入图像的设置下，本文的方法也可以取得与最先进对比方法相当的效果。

表 3 在CityPersons数据集上的消融实验结果(%)

方法	交叉网络	互监督检测器	Stripping-NMS	Reasonable (MR^{-2})	Heavy (MR^{-2})
baseline				11.74	45.24
本文	√		√	11.48	43.64
	√	√		10.76	42.44
	√	√	√	10.48	40.81

表 4 在CrowdHuman数据集上的消融实验结果(%)

方法	新的第一阶段	互监督检测器	Stripping-NMS	MR^{-2}	AP	Recall
baseline				42.73	85.88	80.74
本文	√		√	42.38	88.63	83.04
	√	√		41.88	89.44	83.41
	√	√	√	41.04	91.25	84.26

表 5 不同方法在CityPersons数据集上的性能比较(%)

方法	主干网络	输入尺度	Reasonable(MR^{-2})	Heavy(MR^{-2})
OR-CNN[20]	VGG-16	1×	12.80	55.70
MGAN[21]	VGG-16	1×	11.50	51.70
Adaptive-NMS[9]	VGG-16	1×	12.90	56.40
R^2NMS[10]	VGG-16	1×	11.10	53.30
EGCL[22]	VGG-16	1×	11.50	50.00
RepLoss[6]	ResNet-50	1×	13.20	56.90
CrowDet[23]	ResNet-50	1×	12.10	40.00
文献[24]	ResNet-50	–	11.60	47.30
RepLoss[6]	ResNet-50	1.3×	11.60	55.30
CrowDet[23]	ResNet-50	1.3×	10.70	**38.00**
NOH-NMS[25]	ResNet-50	1.3×	10.80	–
baseline	ResNet-50	1.3×	11.74	45.24
本文方法	ResNet-50	1.3×	**10.48**	40.81

表 6　不同方法在CrowdHuman数据集上的性能比较(%)

方法	主干网络	MR^{-2}	AP	Recall
Faster R-CNN[11]	VGG-16	51.21	85.09	77.24
Adaptive-NMS[9]	VGG-16	49.73	84.71	91.27
JointDet[7]	ResNet-50	46.50	–	–
R^2NMS[10]	ResNet-50	43.35	89.29	93.33
CrowdDet[23]	ResNet-50	41.40	90.70	83.70
DETR[26]	ResNet-50	45.57	89.54	**94.00**
NOH-NMS[25]	ResNet-50	43.90	89.00	92.90
文献[8]	ResNet-50	50.16	87.31	FPN
V2F-Net[27]	ResNet-50	42.28	91.03	84.20
baseline	ResNet-50	42.73	85.88	80.74
本文方法	ResNet-50	**41.04**	**91.25**	84.26
增益	–	–1.69	+5.37	+3.52

表 7　DetNet与Cascade R-CNN结合的性能(%)

方法	主干网络	MR^{-2}	AP	Recall
本文方法	Detnet-59	39.94	91.23	93.05
Cascade R-CNN+本文方法	Detnet-59	38.02	91.75	93.14

为了进一步评估本文所提方法在拥挤场景下行人检测方面的性能，又将本文所提出的模型与其他最先进的行人检测方法在CrowdHuman数据集上进行了比较。表6展示了在CrowdHuman数据集上本文方法与其他先进行人检测方法之间的比较结果。除Faster R-CNN和Adaptive-NMS以外，其余所有方法都采用相同的主干网络(backbone)ResNet-50，并使用相同大小的输入图像进行评估。从表6可以看出，本文所提出的方法优于大多数主流行人检测器。这些实验验证了本文模型的有效性，揭示了本文所提出的模型在拥挤场景下检测遮挡行人方面的优势。

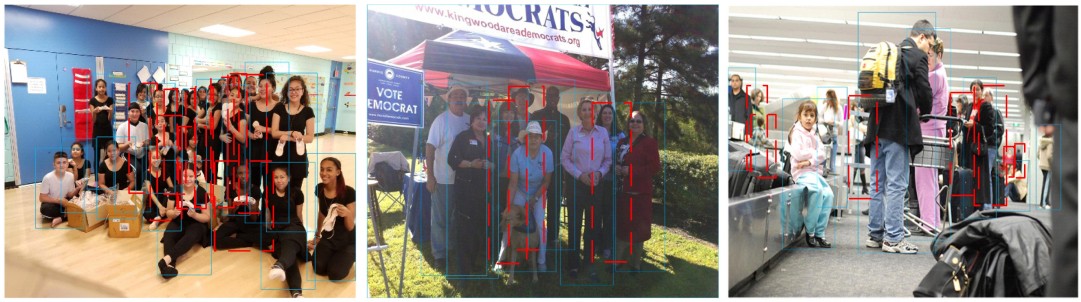
实验发现，以VGG-16或ResNet-50为主干网络的baseline，其性能还有待提升。因此，本文重新选取DetNet[28]作为主干网络并以Cascade R-CNN[29]作为新的baseline进行进一步实验验证。实验结果如表7所示，本文的方法仍然可以在CrowdHuman数据集上提高Cascade R-CNN的性能。

图3展示了本文提出的方法与Faster R-CNN在拥挤场景下的行人检测结果。可视化的分数阈值为0.3，图3(a)是Faster R-CNN算法的检测结果，其中蓝色实线框代表Faster R-CNN检测到的边界框，虚线框代表未检测到或者被NMS抑制而漏检的行人目标。图3(b)是本文提出方法的检测结果。可以直观地看出，Faster R-CNN算法未检测到与漏检的实例在本文所提出的方法中几乎全部可以准确回归。

此外，为了评估本文模型的泛化性，首先基于CrowdHuman数据集的行人头部框标注(h)与全身框标注(f)对模型进行训练，然后在CityPersons数据集进行了测试。如表8所示，本文的方法在Reasonable子集与Heavy子集上分别取得了9.61%与40.23%的MR^{-2}效果。同时又基于CrowdHuman数据集与CityPersons中共有的可见框标注(v)与全身框标注(f)进行了联合训练，然后在CityPersons测试集进行了测试，取得了目前几乎最先进的性能表现，在泛化性能上也优于文献[8]等方法。

4　结束语

针对拥挤场景中行人检测面临的诸多挑战，本文提出了一种新的Anchor free算法和Anchor base双阶段算法之间的平衡策略。首先，通过Anchor free检测头的预检测得到了行人锚框的预设信息与人群区域的密度信息，同时解决了Anchor base算法手工设置锚框的局限性问题和NMS算法中的密度难以判断的问题。然后，提出了一种新的头部-行人互监督检测网络，利用行人头部-全身锚框对的对应关系，通过两次互相监督有效减少了全身检测的漏检及误检情况。最后，设计了一种新的NMS算法，该算法可根据不同的人群密度区域自适应地选择合适的IoU阈值，并且可以通过预先剥离Anchor free检测头中的高置信度检测结果进一步减少NMS处理时的漏检情况。在极具挑战性的CrowdHuman数据集和CityPersons数据集上的良好性能验证了本文方法的有效性。



(a) Faster R-CNN

(b) 本文方法

图 3 在拥挤场景下本文方法与Faster R-CNN的检测结果比较

表 8 本文方法的泛化性能(%)

方法	训练	测试	Reasonable(MR^{-2})	Heavy(MR^{-2})
文献[8]	CrowdHuman	CityPersons	10.10	50.20
本文方法	CrowdHuman(h&f)	CityPersons		40.23
	CrowdHuman+CityPersons(v&f)	CityPersons	8.84	39.27

参 考 文 献

[1] YE Mang, SHEN Jianbing, LIN Gaojie, *et al*. Deep learning for person Re-identification: A survey and outlook[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022, 44(6): 2872–2893. doi: 10.1109/TPAMI.2021.3054775.

[2] MARVASTI-ZADEH S M, CHENG Li, GHANEI-YAKHDAN H, *et al*. Deep learning for visual tracking: A comprehensive survey[J]. *IEEE Transactions on Intelligent Transportation Systems*, 2022, 23(5): 3943–3968. doi: 10.1109/TITS.2020.3046478.

[3] 贲晛烨, 徐森, 王科俊. 行人步态的特征表达及识别综述[J]. 模式识别与人工智能, 2012, 25(1): 71–81. doi: 10.3969/j.issn.1003-6059.2012.01.010.

BEN Xianye, XU Sen, and WANG Kejun. Review on pedestrian gait feature expression and recognition[J]. *Pattern Recognition and Artificial Intelligence*, 2012, 25(1): 71–81. doi: 10.3969/j.issn.1003-6059.2012.01.010.

[4] 邹逸群, 肖志红, 唐夏菲, 等. Anchor-free的尺度自适应行人检测算法[J]. 控制与决策, 2021, 36(2): 295–302. doi: 10.13195/j.kzyjc.2020.0124.

ZOU Yiqun, XIAO Zhihong, TANG Xiafei, *et al*. Anchor-free scale adaptive pedestrian detection algorithm[J]. *Control and Decision*, 2021, 36(2): 295–302. doi: 10.13195/j.kzyjc.2020.0124.

[5] ZHOU Chunluan and YUAN Junsong. Bi-box regression for pedestrian detection and occlusion estimation[C]. The 15th European Conference on Computer Vision, Munich, Germany, 2018: 135–151. doi: 10.1007/978-3-030-01246-5_9.

[6] WANG Xinlong, XIAO Tete, JIANG Yuning, *et al*. Repulsion loss: Detecting pedestrians in a crowd[C]. The 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 7774–7783. doi: 10.1109/CVPR.2018.00811.

[7] CHI Cheng, ZHANG Shifeng, XING Junliang, *et al*. Relational learning for joint head and human detection[C]. The Thirty-Fourth AAAI Conference on Artificial Intelligence, New York, USA, 2020: 10647–10654. doi: 10.1609/aaai.v34i07.6691.

[8] 陈勇, 谢文阳, 刘焕淋, 等. 结合头部和整体信息的多特征融合行人检测[J]. 电子与信息学报, 2022, 44(4): 1453–1460. doi: 10.11999/JEIT210268.

CHEN Yong, XIE Wenyang, LIU Huanlin, *et al*. Multi-feature fusion pedestrian detection combining head and overall information[J]. *Journal of Electronics & Information Technology*, 2022, 44(4): 1453–1460. doi: 10.11999/JEIT210268.

[9] LIU Songtao, HUANG Di, and WANG Yunhong. Adaptive

NMS: Refining pedestrian detection in a crowd[C]. The 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 2019: 6452–6461. doi: 10.1109/CVPR.2019.00662.

[10] HUANG Xin, GE Zheng, JIE Zequn, *et al*. NMS by representative region: Towards crowded pedestrian detection by proposal pairing[C]. The 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2020: 10747–10756. doi: 10.1109/CVPR42600.2020.01076.

[11] REN Shaoqing, HE Kaiming, 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. doi: 10.1109/TPAMI.2016.2577031.

[12] ZHOU Xingyi, WANG Dequan, and KRÄHENBÜHL P. Objects as points[EB/OL]. https://arxiv.org/abs/1904.07850, 2019.

[13] LIN T Y, GOYAL P, GIRSHICK R, *et al*. Focal loss for dense object detection[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020, 42(2): 318–327. doi: 10.1109/TPAMI.2018.2858826.

[14] ZHENG Zhaohui, WANG Ping, REN Dongwei, *et al*. Enhancing geometric factors in model learning and inference for object detection and instance segmentation[J]. *IEEE Transactions on Cybernetics*, 2022, 52(8): 8574–8586. doi: 10.1109/TCYB.2021.3095305.

[15] BODLA N, SINGH B, CHELLAPPA R, *et al*. Soft-NMS--improving object detection with one line of code[C]. The 2017 IEEE International Conference on Computer Vision, Venice, Italy, 2017: 5562–5570. doi: 10.1109/ICCV.2017.593.

[16] SHAO Shuai, ZHAO Zijian, LI Boxun, *et al*. CrowdHuman: A benchmark for detecting human in a crowd[EB/OL]. https://arxiv.org/abs/1805.00123, 2018.

[17] ZHANG Shanshan, BENENSON R, and SCHIELE B. CityPersons: A diverse dataset for pedestrian detection[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017, 4457–4465. doi: 10.1109/CVPR.2017.474.

[18] SHAO Xiaotao, WANG Qing, YANG Wei, *et al*. Multi-scale feature pyramid network: A heavily occluded pedestrian detection network based on ResNet[J]. *Sensors*, 2021, 21(5): 1820. doi: 10.3390/s21051820.

[19] LIN T Y, DOLLÁR P, GIRSHICK R, *et al*. Feature pyramid networks for object detection[C]. The 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 936–944. doi: 10.1109/CVPR.2017.106.

[20] ZHANG Shifeng, WEN Longyin, BIAN Xiaobian, *et al*. Occlusion-aware R-CNN: Detecting pedestrians in a crowd[C]. The 15th European Conference on Computer Vision, Munich, Germany, 2018: 657–674. doi: 10.1007/978-3-030-01219-9_39.

[21] PANG Yanwei, XIE Jin, KHAN M H, *et al*. Mask-guided attention network for occluded pedestrian detection[C]. The 2019 IEEE/CVF International Conference on Computer Vision, Seoul, Korea (South), 2019: 4966–4974. doi: 10.1109/ICCV.2019.00507.

[22] LIN Zebin, PEI Wenjie, CHEN Fanglin, *et al*. Pedestrian detection by exemplar-guided contrastive learning[J]. *IEEE Transactions on Image Processing*, 2023, 32: 2003–2016. doi: 10.1109/TIP.2022.3189803.

[23] CHU Xuangeng, ZHENG Anlin, ZHANG Xiangyu, *et al*. Detection in crowded scenes: One proposal, multiple predictions[C]. The 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2020: 12211–12220. doi: 10.1109/CVPR42600.2020.01223.

[24] 陈勇, 刘曦, 刘焕淋. 基于特征通道和空间联合注意机制的遮挡行人检测方法[J]. 电子与信息学报, 2020, 42(6): 1486–1493. doi: 10.11999/JEIT190606.

CHEN Yong, LIU Xi, and LIU Huanlin. Occluded pedestrian detection based on joint attention mechanism of channel-wise and spatial information[J]. *Journal of Electronics & Information Technology*, 2020, 42(6): 1486–1493. doi: 10.11999/JEIT190606.

[25] ZHOU Penghao, ZHOU Chong, PENG Pai, *et al*. NOH-NMS: Improving pedestrian detection by nearby objects hallucination[C]. The 28th ACM International Conference on Multimedia, Seattle, USA, 2020: 1967–1975. doi: 10.1145/3394171.3413617.

[26] LIN M, LI Chuming, BU Xingyuan, *et al*. DETR for crowd pedestrian detection[EB/OL]. https://arxiv.org/abs/2012.06785, 2020.

[27] SHANG Mingyang, XIANG Dawei, WANG Zhicheng, *et al*. V2F-Net: Explicit decomposition of occluded pedestrian detection[EB/OL]. https://arxiv.org/abs/2104.03106, 2021.

[28] LI Zeming, PENG Chao, YU Gang, *et al*. DetNet: A backbone network for object detection[EB/OL]. https://arxiv.org/abs/1804.06215, 2018.

[29] CAI Zhaowei and VASCONCELOS N. Cascade R-CNN: Delving into high quality object detection[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 6154–6162. doi: 10.1109/CVPR.2018.00644.

谢明鸿：男，博士，高级工程师，研究方向为行人重识别与图像融合等.

康　斌：男，硕士生，研究方向为图像处理与目标检测.

李华锋：男，博士，教授，研究方向为计算机视觉与图像处理.

张亚飞：女，博士，副教授，研究方向为图像处理与模式识别.

责任编辑：马秀强