重庆大学
CHONGQING UNIVERSITY

Tasks Integrated Networks: Joint Detection and Retrieval for Image Search

(任务集成网络:图像搜索的联合检测与检索)

Lei Zhang , Senior Member, IEEE, Zhenwei He , Yi Yang ,
Liang Wang, Fellow, IEEE, and Xinbo Gao , Senior Member, IEEE

汇报人：彭榕光、金童童

目录

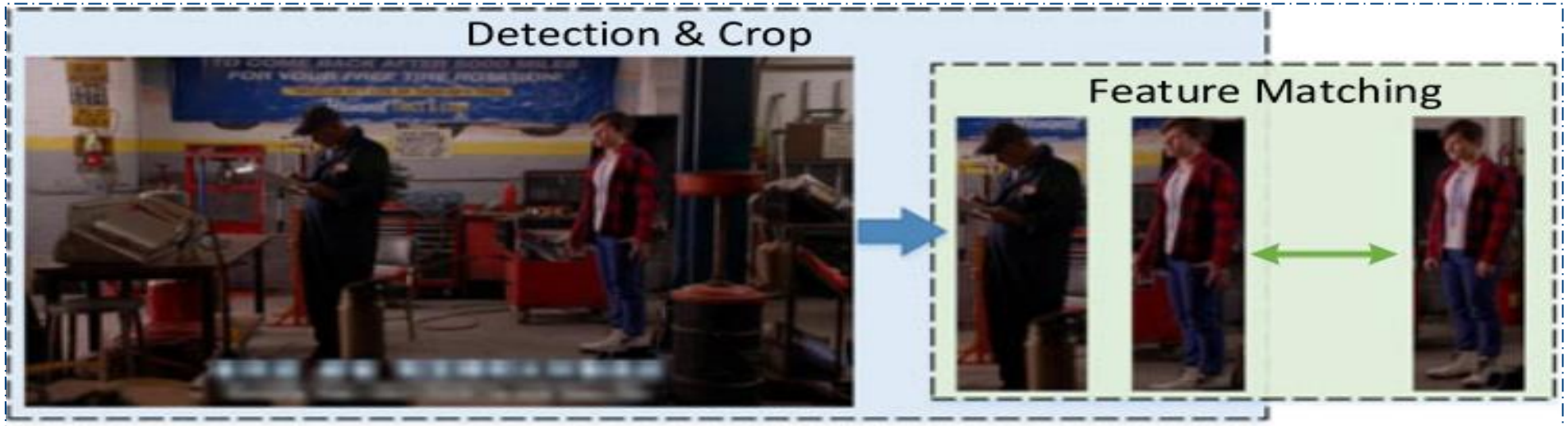
01 研究背景与意义

02 文章相关工作

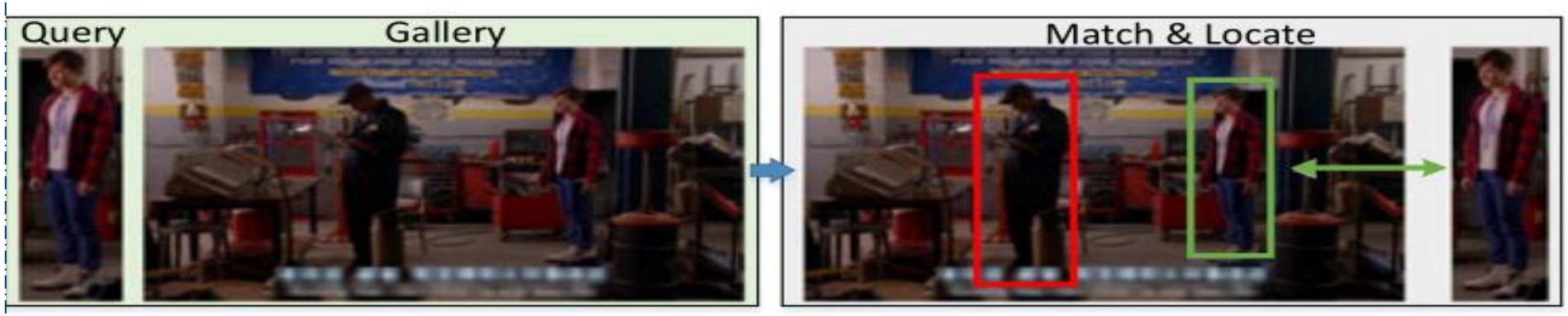
03 实验

一、研究背景与意义

从大型图库图像集中搜索包含感兴趣对象的图像是一个新的但具有挑战性的研究领域。例如，在现实世界的视频监控中，罪犯搜索、多摄像头跟踪等任务与我们的生活息息相关。



(a) Person Re-identification



(b) Person Search (our work)

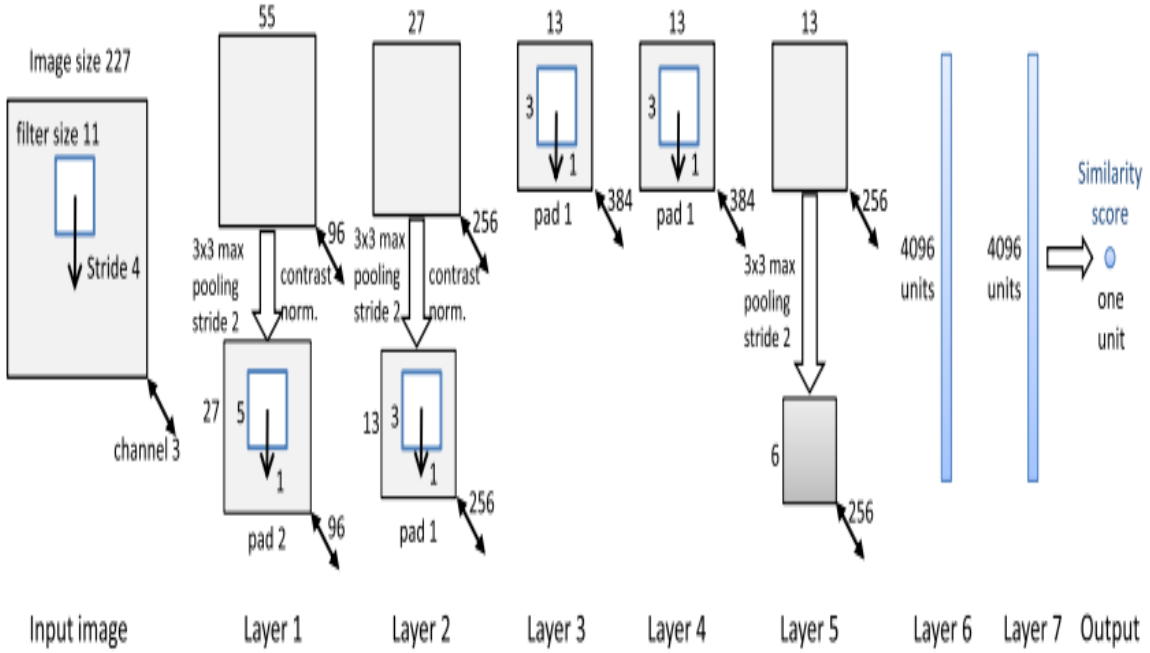
传统的人再识别与人搜索

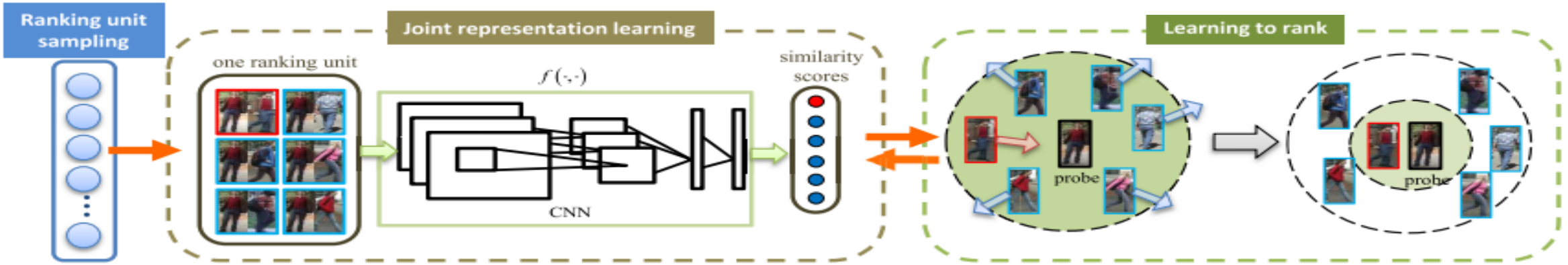
一、现存挑战

- 相机的视角、姿势、遮挡、光照、背景和分辨率不同
- 大多数检测器会出现虚警和不对准的情况
- 在实际应用程序中，这两个独立的子任务对于最终的Re-ID似乎不太友好。

一、国内研究现状

陈世哲[1]等人提出了一种新的人员再识别方法，从探针图像的正确匹配应该定位在整个图库集合的最上面的原则出发。提出了一种有效的排序学习算法，使排序紊乱所带来的代价最小。该方法的深度网络架构如右图,深度排名框架如下图。



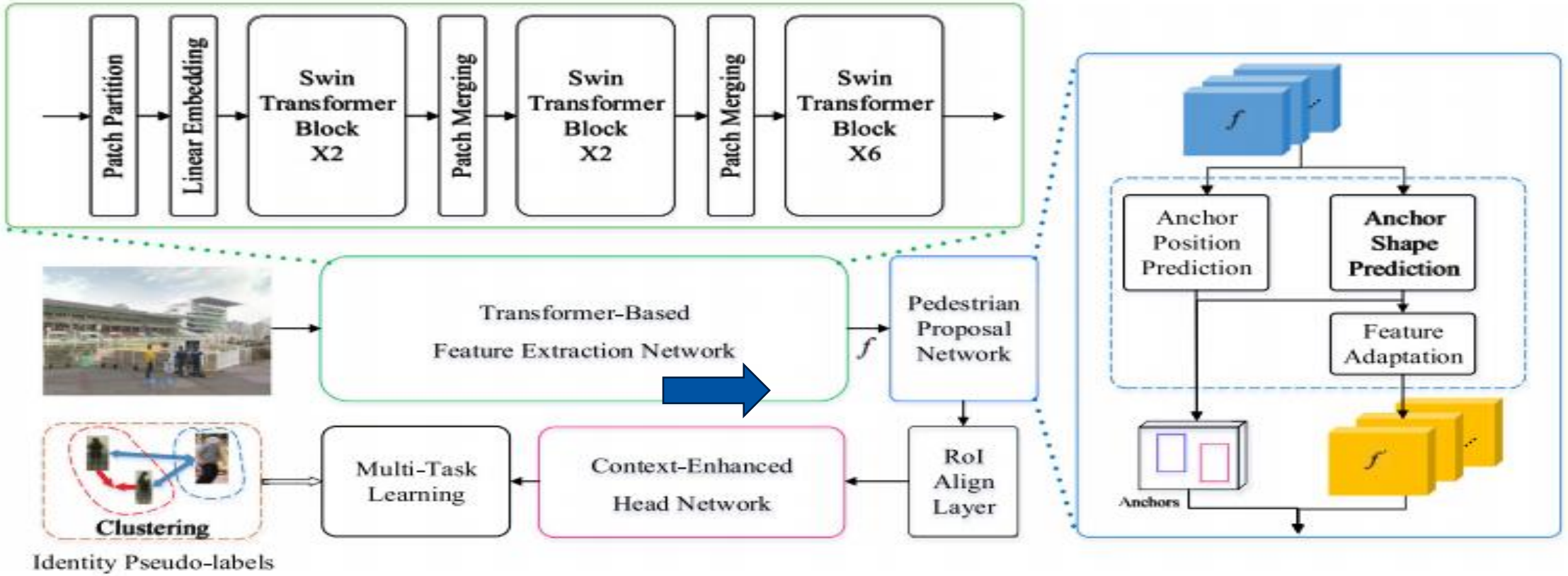


提出的深度排名框架示意图

[1]Imaging Technology; New Imaging Technology Findings from S.Z. Chen and Colleagues Discussed (Deep Ranking for Person Re-Identification via Joint Representation Learning)[J].Journal of Technology Science,2016,

一、国外研究现状

Abdulmutallab El Saddik等人[2]提出了一种用于弱监督人搜索的混合深度网络。混合结构包括基于transformer的特征提取网络和基于全卷积的区域识别头网络。其目的是使模型能够学习不同层次的特征上下文。在该网络中，使用层次视觉变换来提取特征，以获得场景图像的判别表示。提出的弱无监督人员搜索模型的总体结构：如下

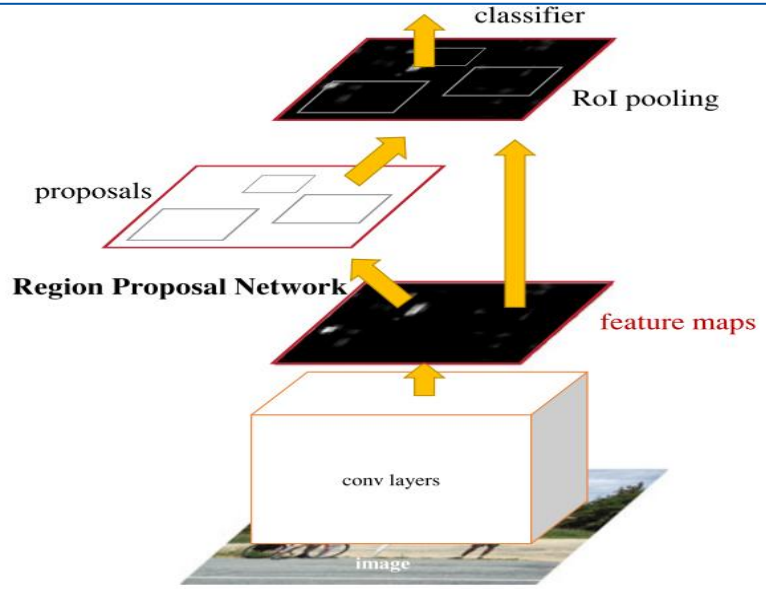


[2] Abdulmutallab El Saddik , et al.Learning feature contexts by transformer and CNN hybrid deep network for weakly supervised person search[J].Computer Vision and Image Understanding,2024,239103906-.

二、文章相关工作

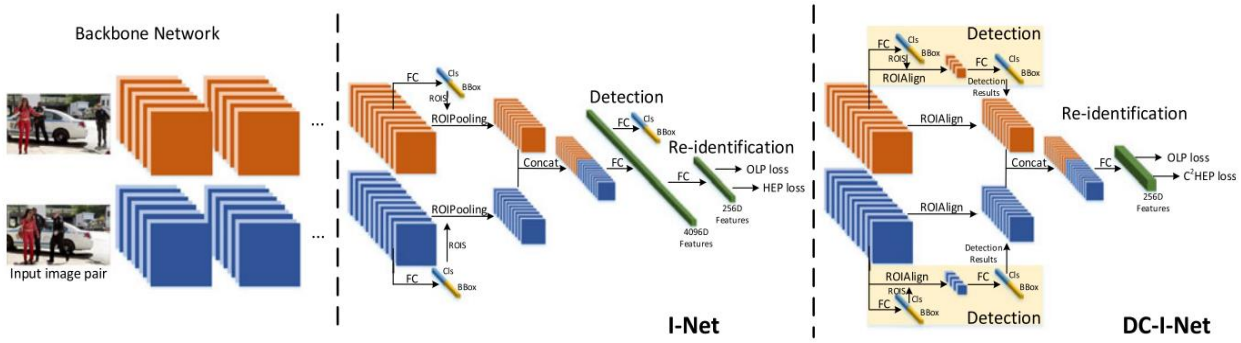
主要解决了一个图像搜索问题，包括人物搜索和纹身搜索。采用Faster-RCNN[3]作为综合图像搜索网络中的目标检测器。

第一阶段生成建议，在第二阶段对目标进行细化，以实现更精确的检测，以端到端方式进行训练，达到最先进的检测性能



Faster-R-CNN是一个单一的、统一的目标检测网络。

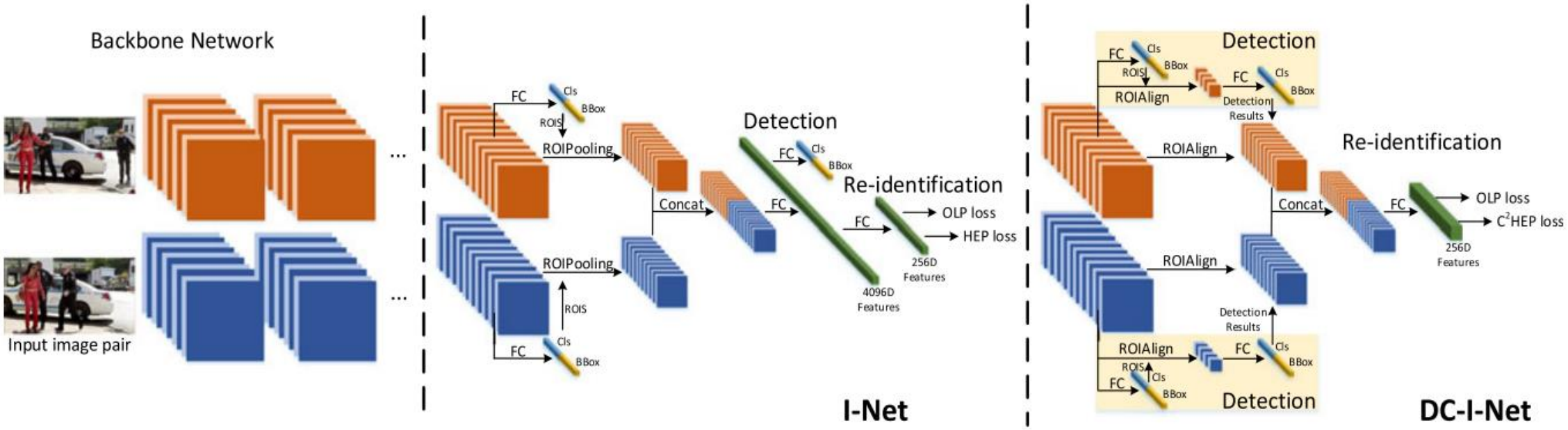
在I-Net网络部分通过产生大量的负样本来限制正对，可以有效地缓解停滞问题



新设计的多任务网络可以将检测损失、度量损失和分类损失相结合，同时进行训练，从而实现更准确、更人性化的人物搜索。

[3]:S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards realtime object detection with region proposal networks," IEEE Trans.Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

二、I-Net网络



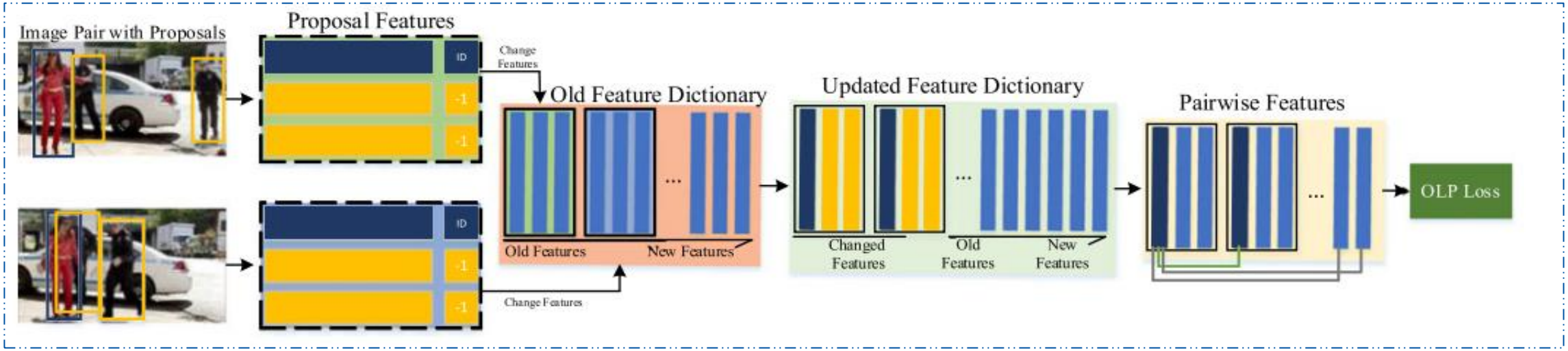
I-Net优点：OLP损失：通过产生大量的负样本来限制正对，可以有效地缓解停滞问题

I-Net缺点：1）因为用于检测的特征集中在前景和背景之间的区分，而用于检索的特征集中在所有前景之间的区分，因此利用同一层的共享特征表示用于检测和检索任务可能会降低图像搜索的性能

2）检索使用对象建议而不是精确对象，这也可能降低检索性能

二、OLP损失

针对OLP损失部署了一个在线特征词典来解决用于训练的样本稀缺的问题。



1)收集每两个输入图像的特征。
2)对于每一对阳性对:或将(p1,p2),p1和p2设置为锚点。特征(n1,n2,...,nk)存储在特征字典中的与锚进行配对以构造负对。
3)使用右边上式计算OLP损失，使用右边下式计算梯度反向传播优化的梯度。
4)存储输入特征，逐步更新特征字典。

$$\mathcal{L}_{OLP} = -\frac{1}{m}\sum_{i=1}^{m}\log\frac{e^{d(\mathbf{x}_a^i,\mathbf{x}_p^i)}}{e^{d(\mathbf{x}_a^i,\mathbf{x}_p^i)}+\sum_{k=1}^{K}e^{d(\mathbf{x}_a^i,\mathbf{x}_{n_k}^i)}},$$

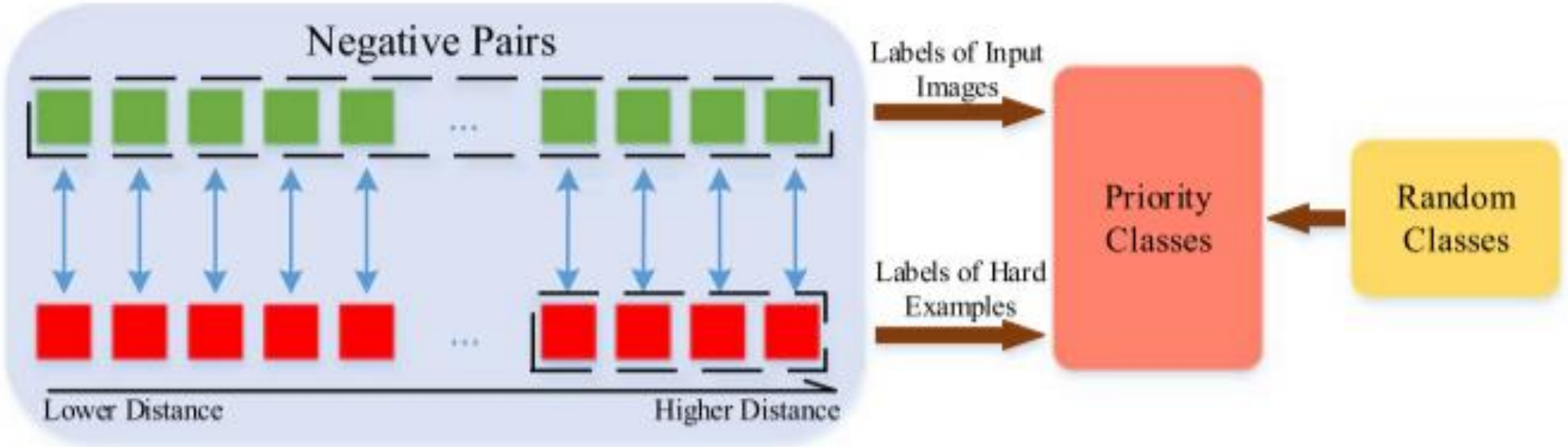
$$\frac{\partial\mathcal{L}_{OLP}}{\partial\mathbf{x}_a^i} = (q^i-1)\mathbf{x}_p^i + \sum_{k=1}^{K}(\hat{q}_k^i\mathbf{x}_{n_k}^i),$$

结构优点：1.I-Net可以通过同时训练度量损失和分类损失来适当匹配给定的两幅图像。

2.以成对图像作为输入，增加了样本数量。这有利于检测器和重新识别器的训练效率和鲁棒性。

二、HEP损失

HEP损失函数旨在回归具有高优先级的身份标签



首先，标记具有身份(即基础事实标签)的人建议(边界框)。其次，选取余弦距离最大的负对作为硬例，记为优先级类。最后，如果优先级类池尚未被填充，则随机选择一些类来填充池，并使用这些类来计算HEP损失。

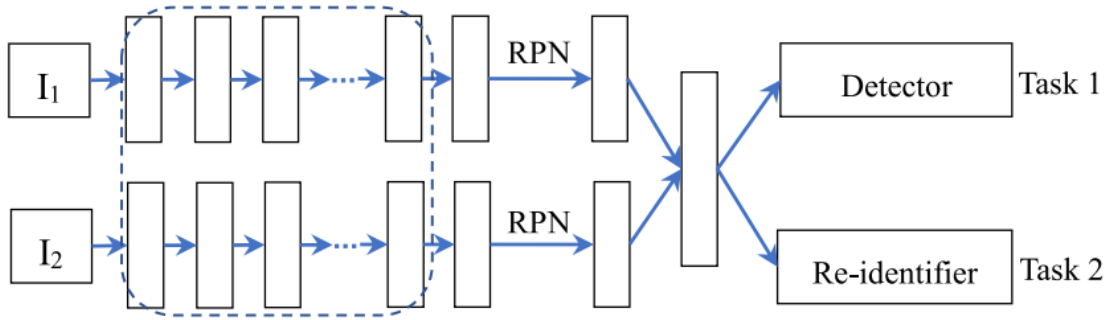
$$\mathcal{L}_{HEP} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j\in\mathcal{P}}\mathbf{1}(label=j)\log\frac{e^{s_j^i}}{\sum_{t=1}^{T}e^{s_t^i}},$$

I-Net是一个端到端的模型，将检测和再识别联合起来进行训练。因此，损耗由检测损耗(LDet)和再识别损耗(LOLP和LHEP)两部分组成，分别表示为:

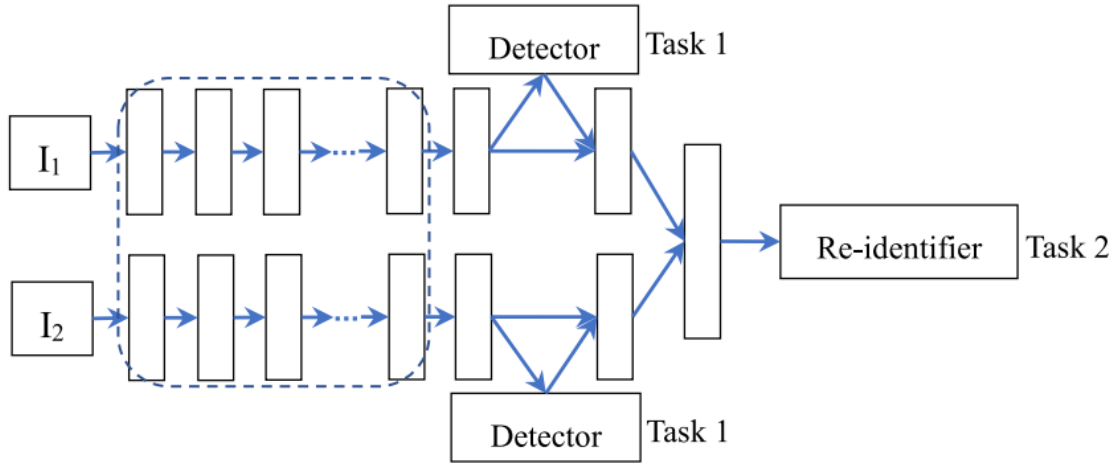
$$\mathcal{L}_{I-Net} = \mathcal{L}_{Det} + \alpha\mathcal{L}_{OLP} + \beta\mathcal{L}_{HEP}$$

二、新的网络结构DC-I-Net

为了解决I-Net在网络架构和训练损失方面的缺陷，提出一种新的网络结构DC-I-Net：



(a) The basic architecture of I-Net



(b) The basic architecture of DC-I-Net

DC-I-Net与I-Net的不同之处在于:

1)将检测器部署在不同层的重标识符前面;

2)使用精炼对象而不是RPN的粗提议来训练重标识符

在此结构下提出了一种类中心引导的硬例优先级损失(C^2HEP)，充分利用更新的输入特征来计算类中心。

二、C²HEP损失

DC-I-Net针对I-Net环境下HEP的难训练问题，提出了类中心引导HEP损失算法(C²HEP)

将余弦相似度输入到HEP的softmax函数中，并构造C²HEP，设计了一个类中心字典，用它们的ground-truth标签来表示：

$$\mathbf{c}_{new}^{j} = \phi \cdot \mathbf{c}_{old}^{j} + (1-\phi) \cdot \mathbf{x}^{(j)}$$

根据softmax函数，我们定义xi属于j类的概率定义为:

$$p_j = \frac{e^{\lambda d(\mathbf{x}_i, \mathbf{c}_j)}}{\sum_{c \in \mathcal{P}} e^{\lambda d(\mathbf{x}_i, \mathbf{c}_c)}}$$

利用负对数似然函数，本文提出的C2HEP损失函数可表示为：

$$\mathcal{L}_{C^2HEP} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j \in \mathcal{P}} \mathbf{1}(label = j) \log \frac{e^{\lambda d(\mathbf{x}_i, \mathbf{c}_j)}}{\sum_{l \in \mathcal{P}} e^{\lambda d(\mathbf{x}_i, \mathbf{c}_l)}}$$

三、实验

■ 数据集

CUHK-SYSU行人搜索数据集：

该数据集从两个来源收集而来：街拍和电影

- 街拍：收集了12,490张图像和6,057个查询人物
- 电影：选取了5,694张图像和2,375个查询人物

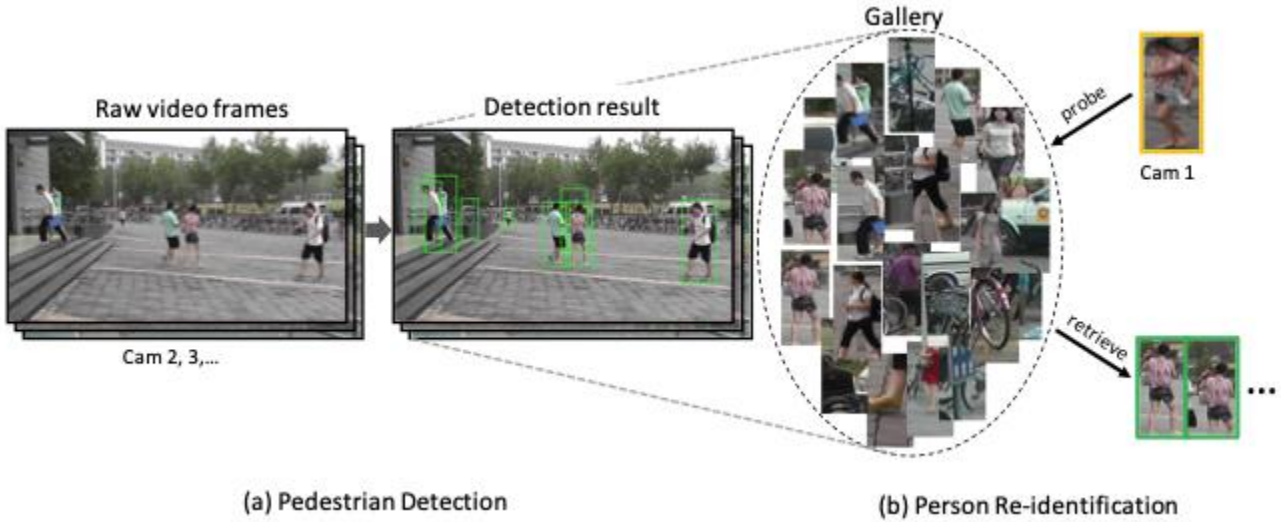
数据被划分为训练集和测试集：

- 训练集包含11,206张图像和5,532个查询人物。
- 测试集包含6,978张图像和2,900个查询人物。
- 训练集和测试集的图像和查询人物没有重叠。

三、实验

■ 数据集

PRW数据集：



- 取自六台摄像机拍摄的10小时视频，其中五台为1080 1920高清，其余一台为576 720标清
- 总共人工标注了11816帧，生成了43110个行人边界框，其中34304个行人被标注了932个ID
- 训练：包含482个标记身份的5134个帧
- 测试：包含450个标记身份的6112个帧

三、实验

■ 数据集

Webtattoo数据集：

数据集由三部分组成：

- 三个小规模(小于10K)纹身数据集的组合
- 从互联网上收集了超过30万张干扰物纹身图像
- 志愿者绘制的300幅纹身素描

纹身检测和图像搜索训练：400个纹身类别的1428张图像

比较不同模型的检测性能：使用了来自200个纹身类别的755幅图像

比较不同模型的搜索性能：使用包含200个图像(每个纹身类别一个图像)的查询集合来从包含355个纹身图像的图库集中搜索图像。

三、实验

■ 环境和参数

环境：

I-Net和DC-I-Net在Caffe和py-faster-rcnn平台上实现，用于模型训练和评估

参数:

- 在I-Net和DC-I-Net中，权衡参数α和β都被设置为1
- 学习率被初始化为0.001，在40k次迭代后下降到0.0001
- 总共设置了70k次迭代以实现收敛

三、实验

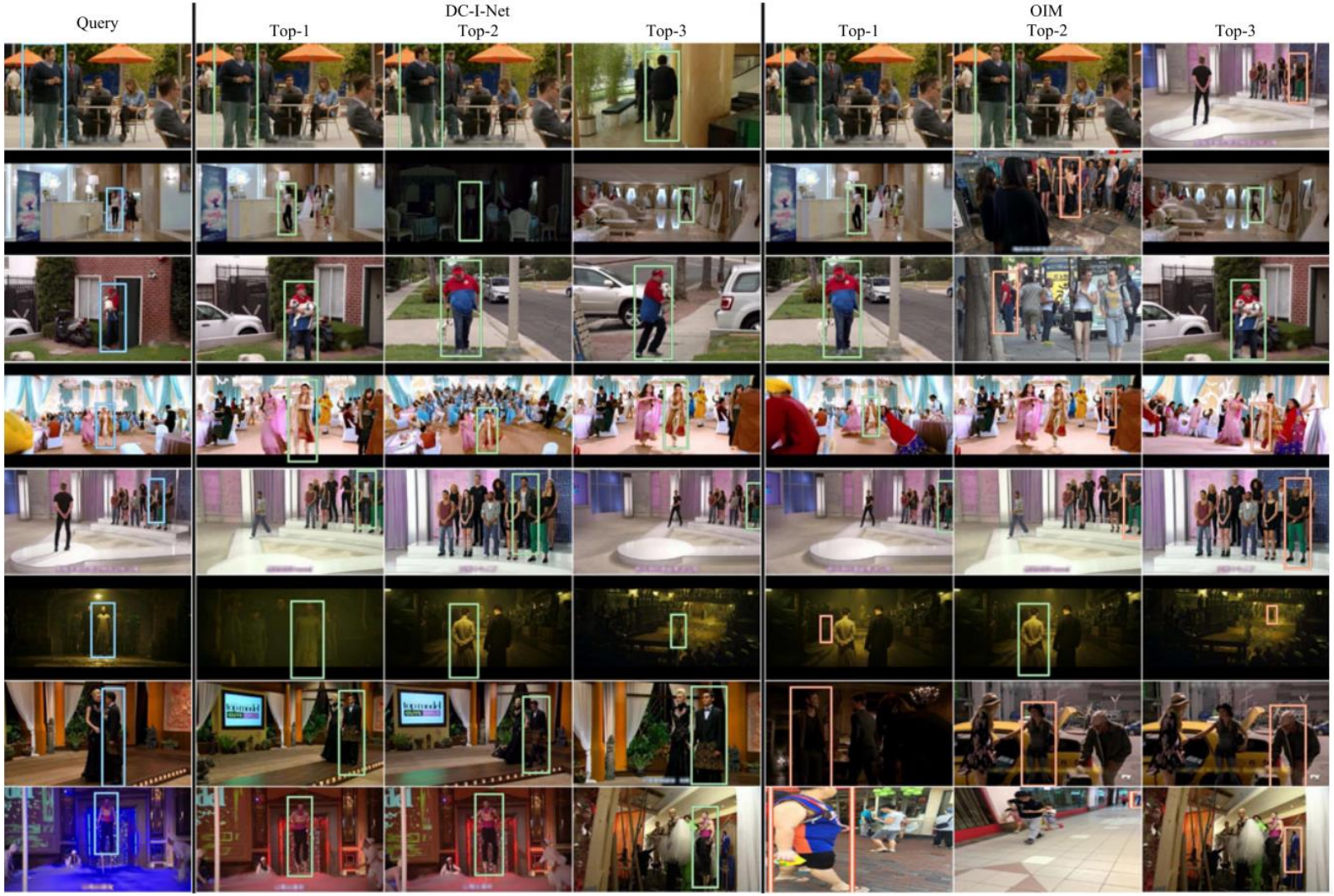
CUHK-SYSU实验

Detector	Re-id Method	mAP(%)	Top-1(%)
ACF	DSIFT [65]+Euclidean	21.7	25.9
	DISFT [65]+KISSME [41]	32.3	38.1
	BOW [66]+KISSME [41]	42.4	48.4
	LOMO [40]+XQDA [40]	55.5	63.1
	IDNet [51]	56.5	63.0
CCF	DSIFT [65]+Euclidean	11.3	11.7
	DISFT [65]+KISSME [41]	13.4	13.9
	BOW [66]+KISSME [41]	26.9	29.3
	LOMO [40]+XQDA [40]	41.2	46.4
	IDNet [51]	50.9	57.1
CNN	DSIFT [65]+Euclidean	34.5	39.4
	DISFT [65]+KISSME [41]	47.8	53.6
	BOW [66]+KISSME [41]	56.9	62.3
	LOMO [40]+XQDA [40]	68.9	74.1
	IDNet [51]	68.6	74.8
GT	DSIFT [65]+Euclidean	41.1	45.9
	DISFT [65]+KISSME [41]	56.2	61.9
	BOW [66]+KISSME [41]	62.5	67.2
	LOMO [40]+XQDA [40]	72.4	76.7
	IDNet [51]	73.1	78.3
End-to-End(Initialized model) [57]		55.7	62.7
OIM [51]		75.5	78.7
IAN [67]		76.3	80.1
NPSM [49]		77.9	81.2
RCAA [68]		79.3	81.3
CNN_v+MGTS [47]		83.0	83.7
I-Net		79.5	81.5
Context Graph [50]		84.1	**86.5**
DC-I-Net(VGG16)		83.7	85.8
DC-I-Net(Resnet50)		**86.2**	**86.5**

- 端到端的人搜索方总是优于单独训练检测器和人重新识别器的传统人搜索方法
- 在行人的地面真实包围盒的情况下，传统的重新识别方法仍然显示出比端到端检测和重新识别方法更差的结果
- 带有Resnet50的DC-I-Net达到了86.5%的TOP-1准确率和86.2%的MAP，并且优于所有比较的方法
- 在VGG16主干上，所提出的DCI-Net在TOP-1准确率和MAP上分别比SOTA端到端人员搜索模型(即I-Net)高4.3%和4.2%

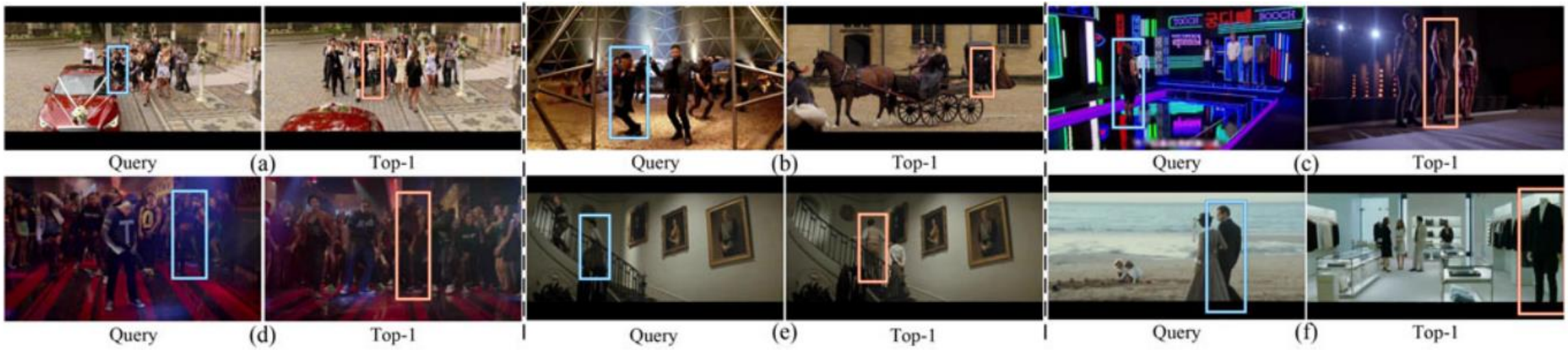
三、实验

■ CUHK-SYSU实验



三、实验

■ CUHK-SYSU实验

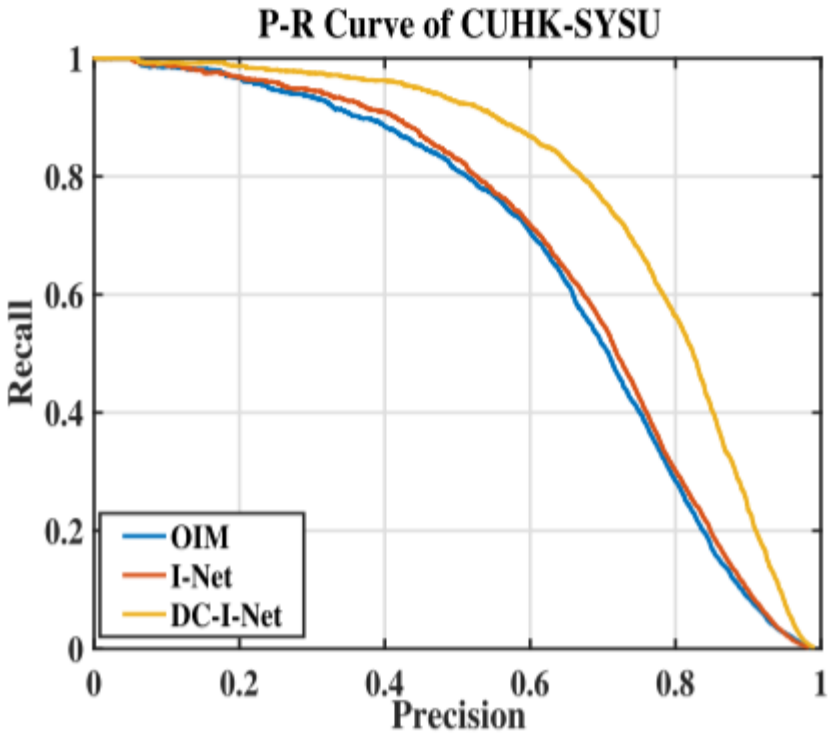


- ✓ 极端条件如下所示。在a和d中，目标人物停留在人群中并且彼此重叠
- ✓ 在图b和图e中，人搜索受到照明、固有因素和姿势变化的影响。
- ✓ 图c、图e和图f中相似的衣服由于相互相似而容易引起错误警报。

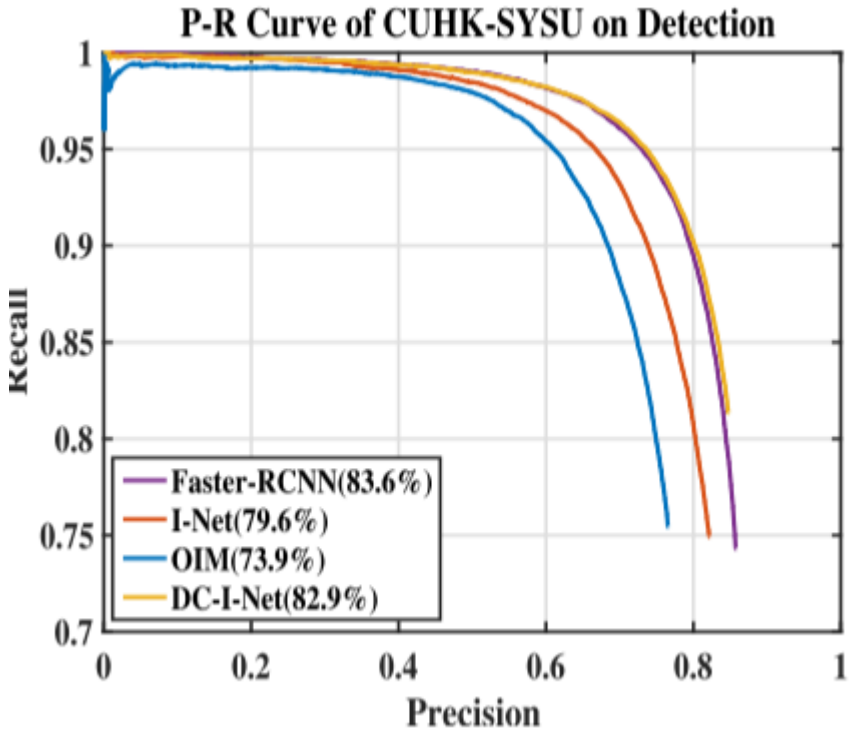
由于人在现实世界的应用中并不总是清晰地呈现，人的搜索任务仍然面临着挑战。

三、实验

■ CUHK-SYSU实验



人员搜索性能



行人检测性能

三、实验

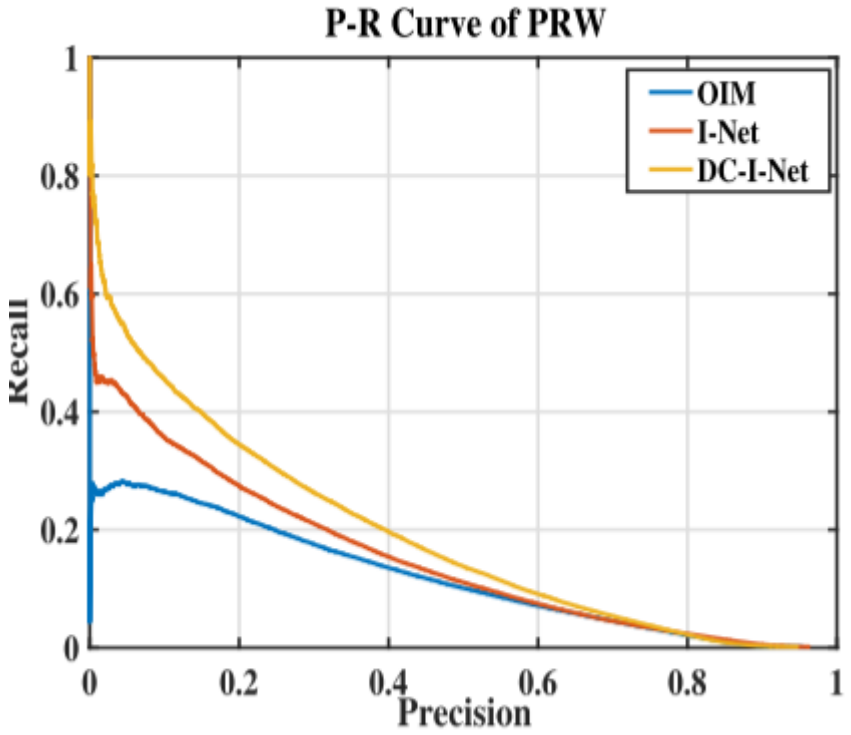
■ PRW实验

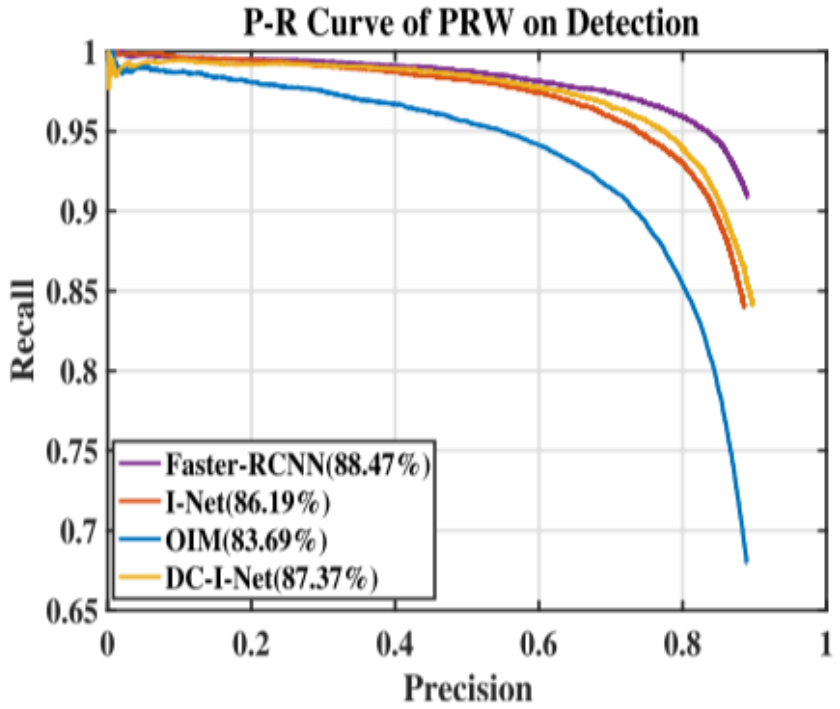
Methods	mAP(%)	Top-1(%)
DPM [69]+BOW [66]	9.7	31.1
DPM [69]+IDE$_{det}$ [45]	18.8	45.9
DPM-Alex+LOMO+XQDA [40]	13.0	34.1
DPM-Alex+IDE$_{det}$ [45]	20.3	47.4
DPM-Alex+IDE$_{det}$ + CWS [45]	20.5	48.3
ACF [18]+LOMO+XQDA [40]	10.5	30.9
ACF [18]+IDE$_{det}$ [45]	17.5	43.8
ACF-Alex+LOMO+XQDA [40]	10.3	30.6
ACF-Alex+IDE$_{det}$ [45]	17.5	43.6
ACF-Alex+IDE$_{det}$ + CWS [45]	17.8	45.2
LDCF [19]+BOW [66]	9.1	29.8
LDCF [19]+LOMO+XQDA [40]	11.0	31.1
LDCF [19]+IDE$_{det}$ [45]	18.3	44.6
LDCF [19]+IDE$_{det}$+CWS [45]	18.3	45.5
OIM [51]	21.3	49.9
NPSM [49]	24.2	53.1
I-Net	25.6	48.7
DC-I-Net(VGG16)	30.4	53.3
DC-I-Net(Resnet50)	**31.8**	**55.1**

- DC-I-Net获得了最好的结果
- 端到端联合训练模型的性能一致优于单独训练的模型，这表明了联合多任务学习在人员搜索中的有效性

三、实验

■ PRW实验

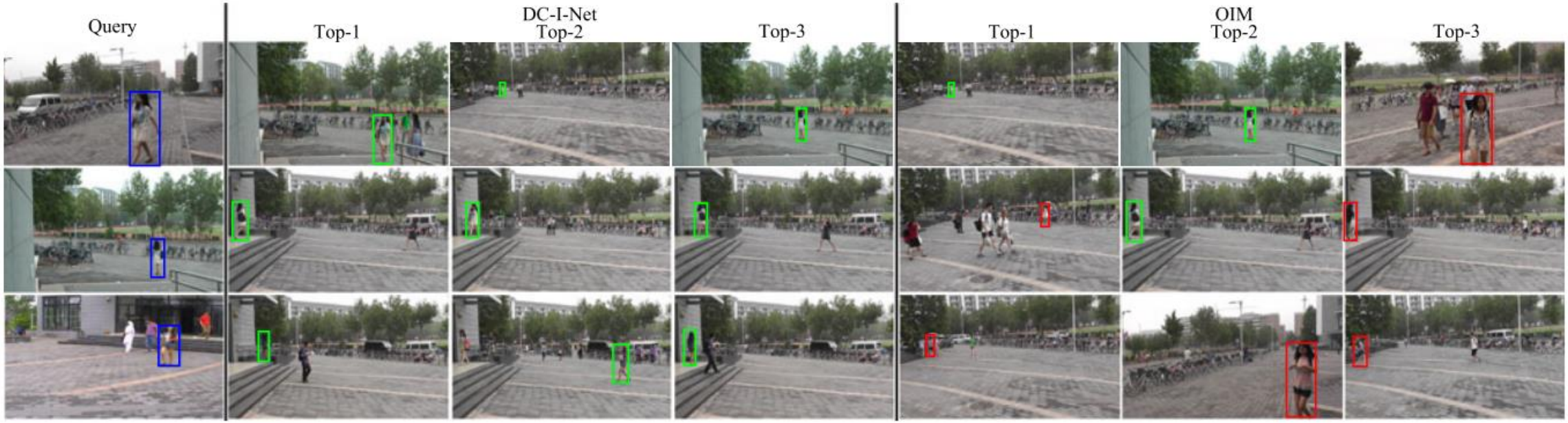




- DC-I-Net在人员搜索任务中表现出最好的性能
- 对于检测任务，我们的模型也接近FasterRCNN单任务检测的最新性能，并优于其他SOTA模型。

三、实验

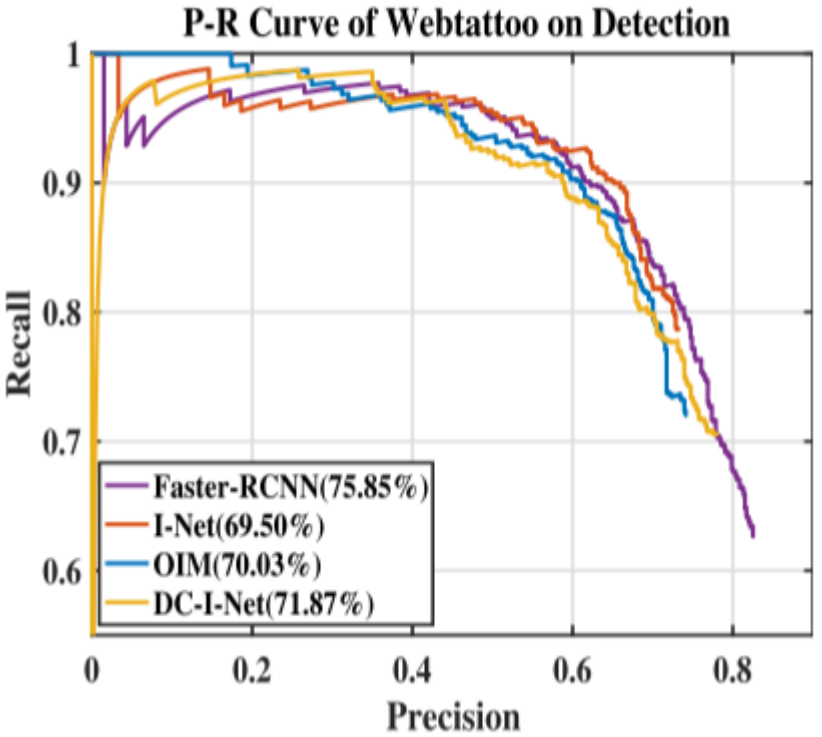
■ PRW实验



- ✓ 每个查询的前3个检索图像中，我们可以看到我们提出的模型可以有比OIM更好的搜索结果。
- ✓ 我们的模型对姿态和分辨率的变化都具有更强的鲁棒性。

三、实验

■ Webtattoo实验

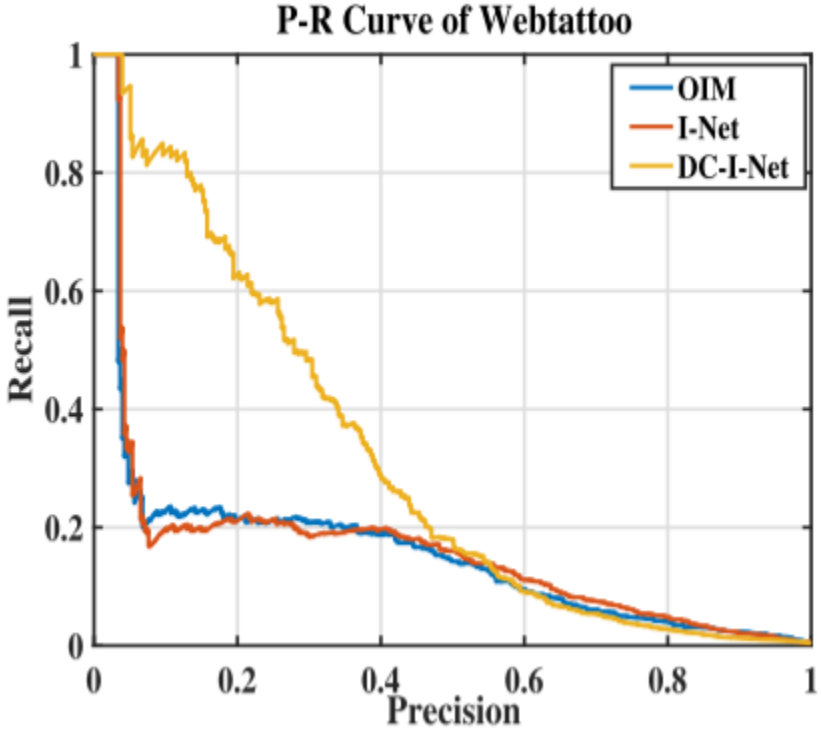


- ✓ 我们模型的检测接近人工标注的地面真实边界框
- ✓ DCI-Net比OIM和I-Net表现出更好的性能，并且它更快地接近于FAST-RCNN检测

三、实验

■ Webtattoo实验

With Detection	mAP(%)	Top-1(%)	Top-5(%)	Top-10(%)
OIM [51]	38.2	39.5	55.5	60.5
I-Net	41.7	44.0	61.0	67.5
DC-I-Net	**48.5**	**51.0**	**66.5**	**71.5**
w/o Detection	mAP(%)	Top-1(%)	Top-5(%)	Top-10(%)
OIM [51]	21.5	23.0	40.0	41.5
I-Net	23.5	25.5	39.5	44.0
DC-I-Net	**30.4**	**31.0**	**51.5**	**64.0**



✓ 检测的模型总是比没有检测部分的模型获得更好的性能，因为检测使模型能够聚焦于图像中的纹身区域，从而使得检测模块的性能大大提高

三、实验

■ Webtattoo实验

三、实验

■ 模型损失的消融研究

➢ 证明了新提出的C2HEP损失比HEP损失的有效性

Loss Type	mAP(%)	Top-1(%)	Top-5(%)	Top-10(%)
I-Net	79.5	81.5	92.2	94.6
I-Net (C^2HEP)	80.9	83.4	94.1	95.2
DC-I-Net (HEP)	81.0	83.0	93.2	95.6
DC-I-Net	**83.7**	**85.8**	**94.3**	**96.1**

✓ I-Net中的HEP损失改为C2HEP损失，MAP和TOP-1的准确率分别提高了1.4%和1.9%

✓ DC-I网络中的C2HEP损耗改为HEP损耗，MAP和TOP-1精度分别降低2.7%和2.8%

三、实验

■ 模型损失的消融研究

➢ 证明OLP损失和基于身份的HEP/C2HEP损失的有效性

Loss Type	mAP(%)	Top-1(%)	Top-5(%)	Top-10(%)
Triplet+HEP	67.8	69.6	87.6	92.2
OLP only	81.3	82.9	93.9	96.0
C^2HEP only	82.2	84.7	94.0	96.0
OLP + HEP	81.9	83.9	93.9	95.6
OLP + C^2HEP	**83.7**	**85.8**	**94.3**	**96.1**

✓ 三元组损失性能较差

✓ 仅使用OLP损耗或仅使用C2HEP损耗的情况下，可以获得更好的性能

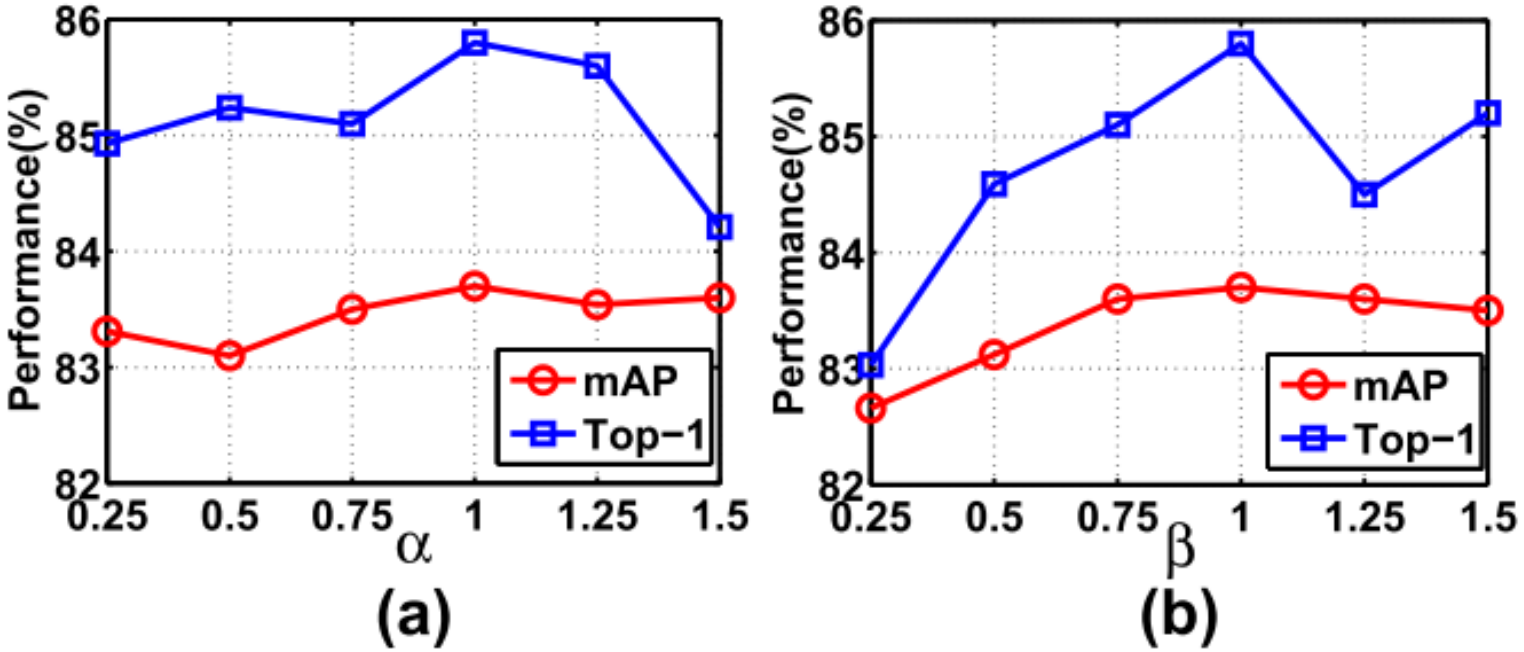
✓ DC-I-Net的性能最好

三、实验

■ 损失权重分析

DC-I-Net的总损耗:

$$L_{DC-I-\mathrm{Net}} = L_{Det} + \alpha L_{OLP} + \beta L_{C^2HEP}$$



三、实验

■ 特征字典大小研究

mAP (%)	20 × 128	40 × 128	60 × 128
OLP only	81.4	81.3	81.8
OLP+C^{2}HEP	82.5	**83.7**	83.1
Top-1 (%)	20 × 128	40 × 128	60 × 128
OLP only	83.5	82.9	83.4
OLP+C^{2}HEP	84.7	**85.8**	85.2

- ✓ 只有OLP损失，增加特征字典的大小并不能提高性能
- ✓ 在OLP和C2HEP损失训练模型的情况下，选择合适大小的特征字典可以获得最好的性能，这是因为C2HEP损失可以探索数据集中所有已标记的身份。

三、实验

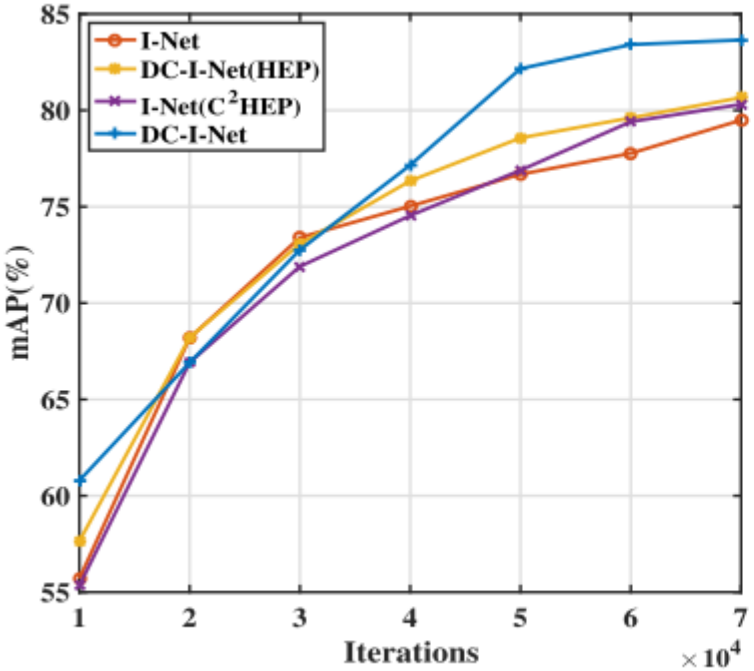
■ 优先级类别数影响

	mAP(%)	Top-1(%)	Top-5(%)	Top-10(%)
50	82.5	84.8	94.1	96.0
100	**83.7**	**85.8**	94.3	**96.1**
1000	83.1	85.0	**94.6**	95.8
5532	82.9	83.7	93.8	95.6

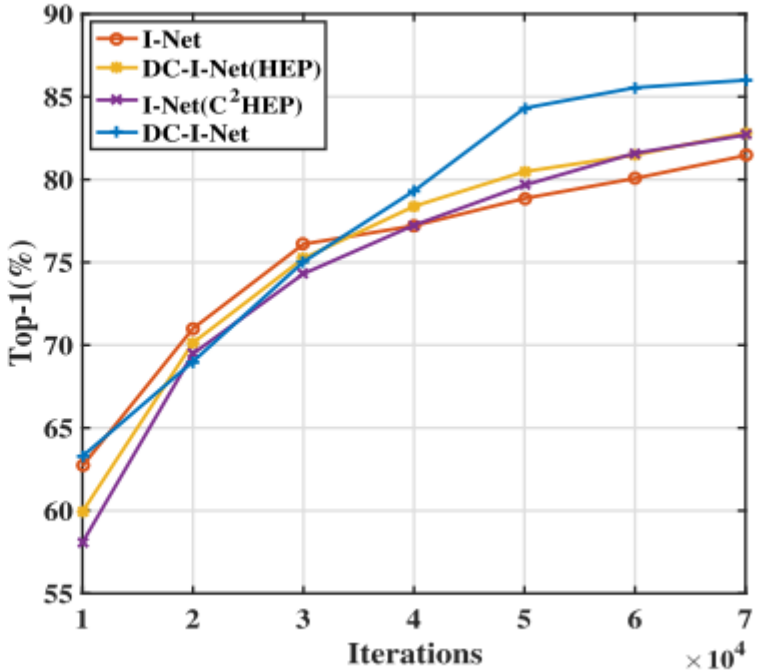
- ✓ 5532表示在不考虑优先级的情况下在损失计算中使用所有被标记的人
- ✓ 100个选择的优先级类别训练的模型好，验证了所提出的硬例优先策略的有效性
- ✓ 随着优先级类数目的增加，性能略有下降，这是因为对真正困难的类的关注可能会减少；如果优先级类的数量太少，模型可能不能很好地探索整个数据集的身份，从而导致性能不佳

三、实验

■ 迭代次数研究



(a) mAP during the Training Phase



(b) Top-1 during the Training Phase

- ✓ MAP和TOP-1精度都随着训练迭代而增加。DC-I-Net呈现出比I-Net更快的上升趋势。
- ✓ 请注意，在训练阶段的40K迭代时，学习速度会降低。

三、实验

■ 输入图像数目研究

	OLP+C²HEP		Contrastive Loss	
	mAP(%)	Top-1(%)	mAP(%)	Top-1(%)
2-input	**86.2**	**86.5**	48.7	45.0
4-input	**85.9**	**87.0**	54.4	54.7
8-input	**85.8**	**85.9**	60.4	60.7

✓ 当输入图像数量较少时，传统的对比损失甚至不起作用

✓ 我们提出的亏损总是运作良好

汇报完毕

恳请大家批评指正

谢谢！

汇报人：彭榕光、金童童