计算机研究与发展
Journal of Computer Research and Development

DOI: 10.7544/issn1000-1239.202330718
x(x): xx–xx, xxxx

虚实结合的行人重识别方法

祁　磊　任子豪　刘俊汐　耿　新
（新一代人工智能技术与交叉应用教育部重点实验室（东南大学）　南京　211189）
（qilei@seu.edu.cn）

Person Re-identification Method Based on Hybrid Real-Synthetic Data

Qi Lei, Ren Zihao, Liu Junxi, and Geng Xin
（*Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications*(*Southeast University*), *Ministry of Education*, *Nanjing* 211189）

Abstract　In recent years, the rapid urbanization and development of the social economy have led to a growing focus on public safety issues. Governments across the world are increasingly promoting the construction of smart cities and intelligent security systems to safeguard the lives and property of citizens and maintain social stability. Person re-identification (ReID) is an essential technology for building smart cities, with significant implications for security monitoring and criminal investigation applications. The goal of person re-identification is to accurately identify specific individuals captured under different cameras. However, due to intra-class differences resulting from various factors such as illumination, viewpoint, occlusion, and pose, person re-identification remains a challenging task in the field of computer vision. Although existing fully supervised person re-identification methods have made significant progress, the scarcity of data and labels poses a bottleneck to further improving model performance. To address this challenge, this paper introduces a more complex and diverse synthetic dataset with easy-to-obtain labels for auxiliary training, and proposes a novel Camera-aware Asymmetric Adversarial Learning (CAAL) method that overcomes intra-class variation among multiple cameras and the domain-shift between real and synthetic data, enabling the learning of camera-invariant feature representations from diverse data sources. Furthermore, to mitigate the impact of misleading information carried by synthetic datasets and prevent the model from overfitting to synthetic data during adversarial training, we propose using an auxiliary network trained on real-world data to constrain the training of the backbone network. We conduct extensive experiments on two public datasets to demonstrate the effectiveness of the proposed method.

Key words　person re-identification; adversarial learning; knowledge distillation; image retrieval; computer vision

摘　要　近年来，随着城市化进程的加速和社会经济的发展，公共安全问题也愈发引起人们的关注. 为了保障社会稳定和公民生命财产安全，各地政府开始大力推进智能安防和智慧城市的建设. 行人重识别就是构建智慧城市的核心技术之一，对安防监控和刑事调查申请具有重要意义. 行人重识别旨在检索不同摄像头下捕捉到的特定人物. 然而，由于光照、视角、遮挡和姿势等造成的类内差异，行人重识别在计算机视觉领域仍然是一项具有挑战性的任务. 受限于数据和标记匮乏，已有的全监督行人重识别任务在模型层面上改进的方法效果基本达到瓶颈. 引入更复杂多样且标记易获得的大型虚拟数据集来进行辅助训练，

收稿日期：2023-09-06；**修回日期**：2023-12-06
基金项目：国家自然科学基金项目（62206052, 62125602, 62076063）；江苏省自然科学基金项目（BK20210224）
This work was supported by the National Natural Science Foundation of China (62206052, 62125602, 62076063) and the Natural Science Foundation of Jiangsu Province (BK20210224).
通信作者：耿新（xgeng@seu.edu.cn）

并提出了一种基于摄像头感知的非对称领域对抗学习方法,同时缓解领域间差异和多摄像头间类内差异的影响,使模型从更丰富多样的数据中学到摄像头差异无关的特征表示.此外,为了缓解虚拟数据集夹带的误导信息带来的不利影响和对抗训练中真实世界数据集的分布向虚拟数据集的数据分布发生偏移的问题,提出使用基于真实数据训练的辅助网络来约束主干网络的训练.实验在2个公开的数据集上进行验证,表明了该方法的有效性.

关键词 行人重识别;对抗学习;知识蒸馏;图像检索;计算机视觉

中图法分类号 TP391

行人重识别任务(person re-identification, ReID)旨在从互不重叠的摄像头收集的图库中匹配与给定查询图像身份相同的行人图像[1-2].近些年以来,随着强大的深度卷积神经网络(convolutional neural network, CNN)的成功应用,全监督场景下的行人重识别任务取得了快速的发展[3-5].但随之而来的是,现有基于模型层面改进的方法在效果上基本达到瓶颈[6-7].为了进一步提升模型性能以及考虑到收集有标记的虚拟数据不需要耗费人力和物力成本,本文提出将大规模虚拟数据集与真实世界数据集结合来训练行人重识别模型的方法,从而克服因真实世界数据匮乏造成的模型过拟合于训练集,而在测试集上表现不佳的问题.然而,虚拟数据集和真实世界数据集往往存在着较大的领域差异,即2个数据分布之间差异较大,如图1所示.直接将虚拟数据和真实数据混合往往会给模型的学习带来困惑并减慢收敛速度,从而无法在真实世界的数据集上取得很好的性能,甚至可能对模型的性能带来损害.

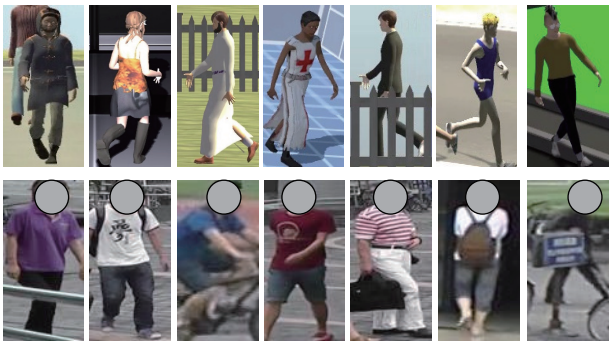


Fig. 1 Illustration of domain discrepancy between virtual dataset and real dataset

图1 虚拟数据集和真实数据集领域差异示意图

除此之外,如图2所示,行人重识别任务的另一大挑战是,在不同摄像头下拍摄出的行人图像通常包含由背景、视角、光照、分辨率、人体姿势等一系列变化引起的显著的类内差异,这些类内差异(intra-class variation)往往会对度量不同摄像头下同一身份的相似性带来不利影响,从而给现有的研究工作带来了巨大的挑战.因此,如何更好地利用大规模的有标记的虚拟数据,使其帮助模型从更丰富多样的数据中学习得到摄像头差异无关的特征表示(camera-invariant feature representations),并提高模型在真实数据上的泛化性能也是本文研究的重点.

Fig. 2 Illustration of intra-class differences of the same person captured under different cameras

图2 同一人在不同摄像头捕捉下的类内差异示意图

针对上述问题,本文提出了基于摄像头感知的对抗迁移学习方法,通过同时建模领域差异和多摄像头间类内差异,使模型从更丰富多样的数据中学习得到更鲁棒的摄像头差异无关的特征表示.具体地说,受 域对抗迁移网络[8](domain-adversarial neural network, DANN)将“对抗机制”引入神经网络训练方法的启发(DANN通过将特征提取器与领域判别器进行对抗训练,使得特征提取器学习到领域无关(domain-invariant)的特征),为了使虚拟数据集的分布尽可能向真实世界数据集靠拢从而消除领域差异,本文将对抗迁移学习引入行人重识别方法.考虑到同一身份一般是由具有不同参数和环境的多个摄像头记录,从而具有显著不同的外观表现.由此造成的摄像头之间的分布差异,也会使得识别同一身份的样本以及优化类内特征的相似性变得更加困难.所以在DANN迷惑领域判别器的基础上,本文提出进一步地使判别器无法分辨特征提取器处理后的数据(图像的特征表示)来自哪个真实领域的摄像头,从而获得摄像头差异无关的特征表示.

此外，为了缓解虚拟数据集夹带的误导信息带来的不利影响，和对抗训练中真实世界数据集的分布向虚拟数据集的数据分布发生偏移的问题，从而提高模型在真实数据集上的测试效果. 本文引入一路只由真实数据集训练的辅助网络，去约束混合数据训练的主网络对真实数据的预测概率. 本文的主要贡献包括 4 个方面：

1)引入少代价的标记虚拟数据集进行辅助训练的虚实结合的行人重识别框架，以此缓解由于数据匮乏导致的全监督行人重识别任务达到瓶颈的问题；

2)提出一种基于摄像头感知的非对称领域对抗模块，可以同时建模虚拟数据集和真实数据集的领域差异和由不同摄像头拍摄带来的类内差异，使模型从更丰富多样的数据中迁移有效的知识到真实场景中和学到摄像头差异无关的特征表示；

3)引入辅助网络进行训练，将知识蒸馏应用到学习框架中，实现参照真实数据分布空间的目标，以缓解虚拟数据集夹带的误导信息带来的不利影响和真实数据集的分布偏移问题；

4)在 2 个公开的行人重识别数据集上进行了实验验证和分析，实验结果表明我们提出的基于知识迁移的虚实行人重识别方法可以有效的从丰富数据中学到摄像头差异无关的特征表示，以提升模型在真实场景下的泛化性能.

1 相关工作

在监控系统中，受背景、视角、光照、分辨率、人体姿势等一系列变化的影响，同一身份的行人图像通常具有显著的类内差异，这些类内差异往往会对度量不同摄像头下同一身份的相似性带来不利影响，从而给现有的研究工作带来了巨大的挑战. 任务发展的初期，研究人员使用人工设计的特征进行行人重识别[9-12]. 近年来，随着深度学习技术的快速发展，基于深度学习的方法在行人重识别任务中占据主导地位，进而得到了广泛的应用和研究[3-7,13]. 本文从 3 个方面介绍本任务的相关工作.

1.1 全监督的行人重识别方法

由于人体特征呈高度结构化，区分相应的身体部位可以有效地确定身份. 许多最近的工作[6,14-16]聚合来自不同身体部位和全局线索的显著特征用于人物重新识别. 其中，基于局部特征的方法[5-6,16]实现了最先进的性能，该方法将输入图像水平分成固定数量的条带，并从这些条带中聚合特征. 但从多个分支聚合特征向量网络结构较为复杂.

此外，区分行人图像的有效特征表示也是行人重识别任务中的一项重要挑战. 为了处理不完美的边界框检测和身体部位错位，许多工作[11,17-19]利用注意力机制来捕捉和聚焦于显示信息性特征(例如，空间位置和人体姿态)并抑制噪声部分(例如，背景和照明). 显著性加权[20-21]是解决这个问题的一种有效方法. 受到注意力模型的启发，Zhao 等人[20]提出了针对行人重识别的部分对齐表示. 在类似的思想下，一些工作[21-22]也表现出了优越的泛化性能，它们将区域注意力选择子网络融合到行人重识别模型中. 为了学习对姿态变化具有鲁棒性的特征表示，姿态引导的注意力方法[23-24]借助姿态估计和人体分割网络将不同的身体部位特征融合在一起，以至于提升模型的性能.

由于摄像头标记信息在行人重识别任务中较为容易获得，且不同摄像头下同一身份的行人图像通常受类内差异的影响难以进行相似度度量. 近年来，一些研究[25-26]通过摄像头标记信息学习摄像头风格的变化，来解决跨摄像头检索任务中的图像风格变化产生类内差异的问题. 例如，Zhong 等人[25]提出了一种平滑相机风格差异的方法. 利用 CycleGAN 将训练图像的风格样式转换到每个摄像头的风格上，生成后的数据样本与原始训练样本一起形成增强的训练集. 此外，Zhuang 等人[26]提出了一种基于摄像头的批量归一化方法，用于解决行人重识别中的分布差异问题. 在训练中通过对每个摄像头进行归一化，实现映射所有摄像头下的数据到共享空间，从而使模型更好地适应新的摄像头.

为了进一步提高检索精度，一些研究工作还采用了重新排序策略[27-28]和使用特定人物属性进行推断[29]. 一些常用的数据增强方法，如随机擦除[30]和 Cutout[31]也被广泛使用. 值得注意的是，上述策略可以与我们的方法结合使用.

1.2 对抗迁移学习的相关研究

对抗迁移学习中一类很重要的算法就是基于生成对抗的. 随着生成对抗网络(generative adversarial networks, GAN)的发展，已经提出了许多基于生成对抗的方法[32-38]被应用到有监督和无监督行人重识别任务上.Liu 等人[32]考虑了由人体姿态变化引起的重新识别鲁棒性挑战，并提出了一种姿态可转移框架，利用姿态转移样本增强姿态鲁棒性. Qian 等人[33]引入人体姿态用于合成具有不同姿态的人物图像，通过合成姿态规范化的图像，使模型所学到的特征表

示不受姿态变化的影响.Wei 等人[34] 通过尽可能保持前景不变的条件下对不同数据集的背景进行转换,利用生成对抗网络将源域的图像风格转移到目标域中.

上述方法都是在图像层级生成图像,这些方法无法进行端到端实现,在生成的过程中也很难保证生成图像中有关行人身份的内容特征不被破坏或丢失.为此,Zheng 等人[35] 设计了一个融合特征学习和数据生成端到端的联合学习框架,使身份判别模块和图像生成模块相互促进从而生成更可靠的数据.由于该方法仍然需要生成新的图像,其相对于特征学习需要增加 GAN,因此在训练过程中需要耗费更多的资源.

另一类方法是基于数据分布自适应的对抗迁移算法,其在特征层级执行对抗训练,我们的方法属于这一类.区别于生成对抗的迁移学习方法,基于数据分布自适应的对抗迁移算法通常为端到端的训练,通过特征提取器和领域判别器的相互对抗训练,以此学到领域不变的特征.Yu 等人[36] 在此基础上将动态分布自适应融入了对抗学习中,提出了一个动态对抗自适应网络,通过动态调整边界和条件分布关系使模型预测更加稳健准确.Qi 等人[37] 受文献 [8] 的启发,提出了一种基于摄像头感知的域适应方法,以此获取摄像头无关的特征表示.不同于存在的对称对抗方法,我们引入了虚拟数据,Wang 等人[39] 针对标记数据匮乏导致模型泛化能力差的问题,使用了虚拟数据集得到多样化数据,提出了包含动态训练数据集生成和域不变特征学习部分的算法.而我们提出了基于摄像头感知的非对称领域对抗模块来实现从虚拟数据中迁移信息到真实场景中,解决了虚拟数据集和真实数据集存在领域差异的问题.

1.3 知识蒸馏的行人重识别方法

知识蒸馏是模型压缩的一种常用的方法,知识蒸馏是通过构建一个轻量化的学生模型,利用性能更好的教师模型的监督信息来训练这个学生模型,以达到更好的性能和精度.目前,一些行人重识别方法也用到了知识蒸馏[3,40].Zhang 等人[40] 提出了深度相互学习策略,使学生在整个训练过程中协作学习并相互教授.Zhang 等人[3] 提出利用局部特征学习影响全局特征学习,并指出局部特征的学习方法是通过计算局部特征间的最短距离.为了缓解我们方法的对抗过程中,真实世界数据集的分布向虚拟数据集的数据分布偏移的现象,我们引入了知识蒸馏的方法,将只由真实世界数据集训练的辅助网络作为教师网络,约束主模型中目标域身份分类的分支概率,保证真实数据空间不发生偏移,以实现在原始空间基础上更好地利用虚拟数据提升模型在真实数据上的泛化性能.

2 形式化定义

在引入虚拟数据集的行人重识别研究中,我们将给定 2 个数据集X_s和X_t混合进行训练,X_s为虚拟数据集,其定义为$X_s=\left\{\left(\boldsymbol{x}_i^s,\boldsymbol{y}_i^s\right)\right\}_{i=1}^{N_s}$,其中$\boldsymbol{x}_i^s$和$\boldsymbol{y}_i^s$分别为虚拟数据集中第$i$个训练样本和它对应的行人身份标记,$N_s$为虚拟数据集的样本数量.给定一个真实世界数据集,我们将其分为训练集X_t,查询集Q_t和图库G_t.同样的我们将具有N_t个样本的真实世界数据集定义为$X_t=\left\{\left(\boldsymbol{x}_i^t,\boldsymbol{y}_i^t,c_i^t\right)\right\}_{i=1}^{N_t}$,其中$\boldsymbol{x}_i^t$和$\boldsymbol{y}_i^t$分别为真实世界数据集中第$i$个训练样本和它对应的行人身份标记,而$c_i^t$则是真实世界数据集中第$i$个训练样本对应的摄像头标记.我们旨在将$X_s$和$X_t$混合并训练出一个行人重识别模型$F(\cdot|\boldsymbol{\theta})$,对于任何一个查询人物的样本$\boldsymbol{q}$,$\boldsymbol{q}\in Q_t$,行人重识别模型应该生成一个特征向量$\boldsymbol{f}_q$用于在图库$G_t$中查询与$\boldsymbol{q}$有相同身份标记的样本$\boldsymbol{g}$,保证$\boldsymbol{q}$与样本$\boldsymbol{g}$的相似度要比图库中其他样本的相似度要高,如

$$\boldsymbol{g}^*=\underset{\boldsymbol{g}\in G_t}{\operatorname{argmax}}\, sim\left(\boldsymbol{f}_g,\boldsymbol{f}_q\right), \tag{1}$$

其中$\boldsymbol{f}=F(\cdot|\boldsymbol{\theta})\in\mathbb{R}^d$是由行人重识别模型提取的一个$d$维特征向量.$sim(\cdot)$是特征之间的相似度.

3 基于知识迁移的虚实结合行人重识别方法

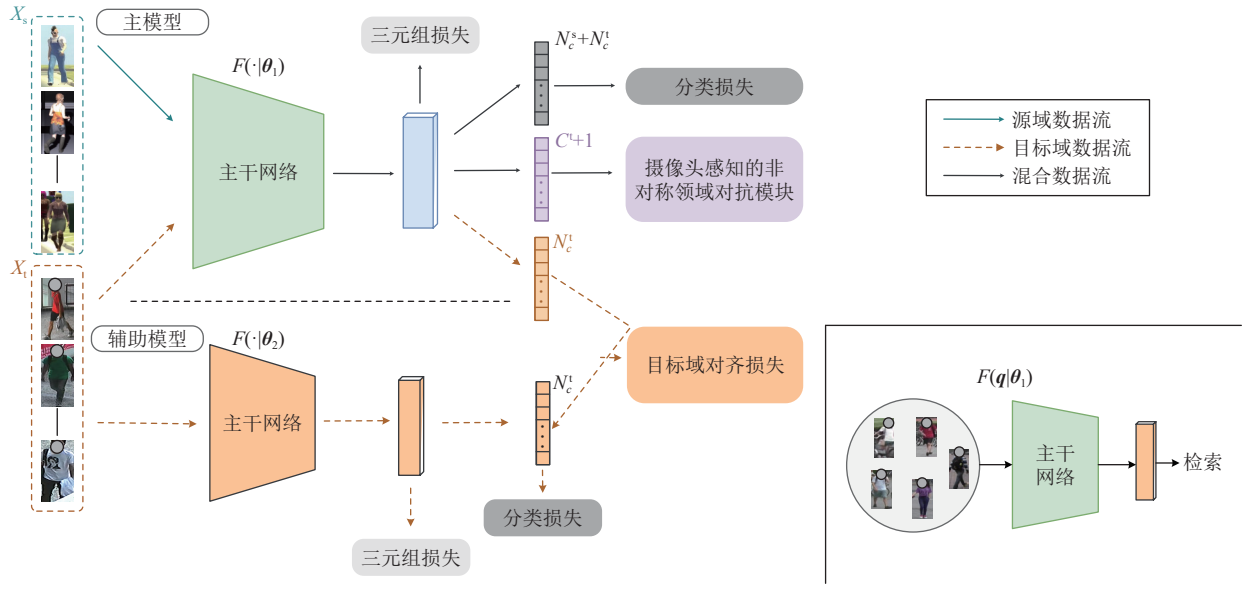
3.1 整体结构

本文提出了一种新型的基于知识迁移的虚实结合行人重识别方法框架,同时缓解领域间差异和多摄像头带来类内差异的影响,使模型从更丰富多样的数据中学到摄像头差异无关的特征表示.整体结构如图 3(a)所示.框架主要包括 3 个模块:

1)源域(虚拟数据)与目标域(真实数据)混合训练的通用全监督行人重识别模型,并增加了一路单独对目标域数据身份分类的分支;

2)在混合数据训练的行人重识别模型上添加了一个摄像头判别器,与主干网络组成基于摄像头感知的非对称领域对抗学习模块;

3)在每一次迭代中,使用混合数据中的目标域样本单独训练的通用全监督行人重识别模型,并使



(a) 方法整体结构 (b) 推理阶段

Fig. 3 The overall framework of the proposed method

图 3 基于知识迁移的虚实结合行人重识别方法框架

用其对模块 1 中的目标域分类分支进行约束对齐.

模块 1 和模块 2 组成了主模型，模块 3 为辅助模型. 模块 2 中主干网络作为特征提取器，与增加的摄像头判别器以对抗的方式迭代学习，其中摄像头判别器要不断的通过学习正确的摄像头标记信息来保证自己判别的准确性，而特征提取器需要生成出可以迷惑摄像头判别器的特征向量以此实现消除领域差异与领域内由摄像头带来的类内差异的问题.

辅助模型则作为教师网络，将由目标域数据单独训练的行人重识别模型输出的身份类别概率作为软标记，以此约束主模型中目标域身份分类的分支概率，减少目标域数据在对抗中发生的数据分布偏移现象. 最后在训练完成后，仅采用主模型中的主干网络作为特征提取器进行推理(见图 3(b))

3.2 通用的行人重识别模型

当前基于 CNN 的全监督行人重识别模型都遵循一个相似的网络结构，通过优化三元组损失[41] 和软分类损失来进行模型训练，身份损失为：

$$\mathcal{L}_{\mathrm{id}} = \mathcal{L}_{\mathrm{scls}} + \mathcal{L}_{\mathrm{tri}}, \tag{2}$$

对于一个批次的样本，通用的行人重识别模型首先对训练数据集进行$P\times K$采样，即在一个小批次中随机对数据集采样P个不同身份的样本，每个身份采样不同的K张. 在本文中，针对混合数据的采样策略调整为均衡采样策略，即在每个小批次中虚拟数据和真实数据的样本数量是一致的，前一半批次为虚拟数据样本，后一半批次为真实数据样本. 若一个小批次包含$P\times K$个样本，那么虚拟数据集和真实世界数据集分别采样$(P/2)\times K$张样本.

采样后的数据进行简单的数据增强后输入到主干网络进行特征提取，该特征提取后会在小批次内计算难三元组损失，从而令小批次中每个样本拉近与它们最不相似的正样本的距离，同时推远与它们最相似负样本的距离. 三元组损失定义为：

$$\mathcal{L}_{\mathrm{tri}} = \frac{1}{N_{\mathrm{b}}}\sum_{i=1}^{N_{\mathrm{b}}} \max\left(0, \|F(\boldsymbol{x}_i\,|\,\boldsymbol{\theta}) - F\left(\boldsymbol{x}_{i,\mathrm{p}}\,|\,\boldsymbol{\theta}\right)\| + a - \|F(\boldsymbol{x}_i\,|\,\boldsymbol{\theta}) - F\left(\boldsymbol{x}_{i,\mathrm{n}}\,|\,\boldsymbol{\theta}\right)\|\right), \tag{3}$$

其中$\boldsymbol{x}_{i,\mathrm{p}}$表示在一个批次中与$\boldsymbol{x}_i$拥有相同身份标记的样本中最不相似的样本，$\boldsymbol{x}_{i,\mathrm{n}}$则表示与$\boldsymbol{x}_i$拥有不同身份标记的样本中最相似的样本. a是一个三元组距离阈值参数，其迫使模型学习，使锚点$\boldsymbol{x}_i$和负例$\boldsymbol{x}_{i,\mathrm{n}}$的距离更大，同时让锚点$\boldsymbol{x}_i$和正例$\boldsymbol{x}_{i,\mathrm{p}}$的距离更小. $\|\cdot\|$代表 L2 范数距离.

最后将特征输入到身份分类器中，每一个样本都会得到它对应所有身份类别的概率，利用分类损失可以鼓励模型学习针对每个身份类别更具有判别性的特征向量. 分类损失定义为：

$$\mathcal{L}_{\mathrm{cls}} = -\frac{1}{N_{\mathrm{b}}}\sum_{i=1}^{N_{\mathrm{b}}} \log p(\boldsymbol{y}_i\,|\,\boldsymbol{x}_i), \tag{4}$$

其中N_{b}，i分别表示一个批次中样本的数量和样本的索引号，$\boldsymbol{x}_i$，$\boldsymbol{y}_i$表示一个批次中第i个样本和它所属于的身份标记. $p(\boldsymbol{y}_i\,|\,\boldsymbol{x}_i)$表示样本$\boldsymbol{x}_i$属于身份标记$\boldsymbol{y}_i$的预测概率.

然而,测试集的行人身份类别从未出现在训练集中.因此,防止行人重识别模型过度拟合训练的身份类别就显得非常重要.文献[42]中提出的标记平滑(label smoothing, LS)是一种广泛使用的方法,可以被用于防止分类任务的过拟合现象.标记平滑后的软分类损失计算更新为:

$$\mathcal{L}_{\text{scls}} = -\frac{1}{N_{\text{b}}}\sum_{i=1}^{N_{\text{b}}}\sum_{j=1}^{N_{\text{c}}} q_j \log p\left(\boldsymbol{y}_j \mid \boldsymbol{x}_i\right), \tag{5}$$

其中

$$q_j = \begin{cases} 1-\dfrac{N_{\text{c}}-1}{N_{\text{c}}}\varepsilon, & \text{if } \boldsymbol{y}_j = \boldsymbol{y}_i, \\ \varepsilon/N_{\text{c}}, & \text{otherwise}, \end{cases}$$

N_{c}为训练集的身份类别个数,j为训练集中的某一身份类别索引,ε为一个小的正常数迫使模型降低对训练集的信任,以此减少过拟合现象.

3.3 基于摄像头感知的非对称领域对抗模块

在行人重识别中,从属于不同摄像头的样本由于背景、视角、光照、人体姿势等一系列变化引起显着的类内差异,这同样也可以被映射在跨领域差异问题上.若将训练集的样本按照从属的摄像头划分,每一个摄像头可以被视作一个子领域.仅仅减少源域和目标域的整体差异并不能有效处理跨摄像头级的差异,这在一定程度上会对学习特征表示的质量产生不利影响.为此,本文提出了一种基于摄像头感知领域适应模块,将从属于不同摄像头的样本映射到共享子空间中.为实现这一目标,提出一种基于摄像头感知的非对称领域对抗学习(camera-aware asymmetric adversarial learning, CAAL),训练流程如图4所示.

传统的对抗性学习[8]通常只处理2个域,通过领域之间的对抗学到领域不变信息.与其不同的是,本文提出的CAAL方法不仅将源域样本映射到目标域空间来降低领域差异,并且可以处理多个子域(即源域和目标域中的摄像头)进一步降低由摄像头引起的类内差异.

假定$X_{\text{s}} = \left\{\left(\boldsymbol{x}_i^{\text{s}}, \boldsymbol{y}_i^{\text{s}}\right)\right\}_{i=1}^{N_{\text{s}}}$和$X_{\text{t}} = \left\{\left(\boldsymbol{x}_i^{\text{t}}, \boldsymbol{y}_i^{\text{t}}, c_i^{\text{t}}\right)\right\}_{i=1}^{N_{\text{t}}}$为源域和目标域中的训练数据,混合训练数据$X = [X_{\text{s}}, X_{\text{t}}]$.如图4所示,在本任务中将源域看作一个摄像头,给定每个源域训练样本的摄像头标记都为"0",c_i^{t}是真实世界数据集中第i个训练样本对应的摄像头标记$c_i^{\text{t}} \in \{0,1,\cdots,C_{\text{t}}-1\}$,混合训练中真实世界数据集的摄像头标记重标记为$c_i^{\text{t}}+1 \in \{1,2,\cdots,C_{\text{t}}\}$,$C_{\text{t}}$为目标域中摄像头的总个数,$C = C_{\text{t}}+1$.对抗学习框架中生成器和判别器以对抗的方式迭代学习,本文跟随DANN[8]的工作,将特征提取器$F(\cdot|\boldsymbol{\theta}_{\text{f}})$视为生成器,另外本文构建了可学习的摄像头判别器$D: \boldsymbol{f} \to \{0,1,\cdots,C_{\text{t}}\}$通过在$C$个摄像头类别上定义的交叉熵损失进行优化.下面将介绍CAAL具体过程.

首先,为了获取更可信的摄像头判别器,固定特

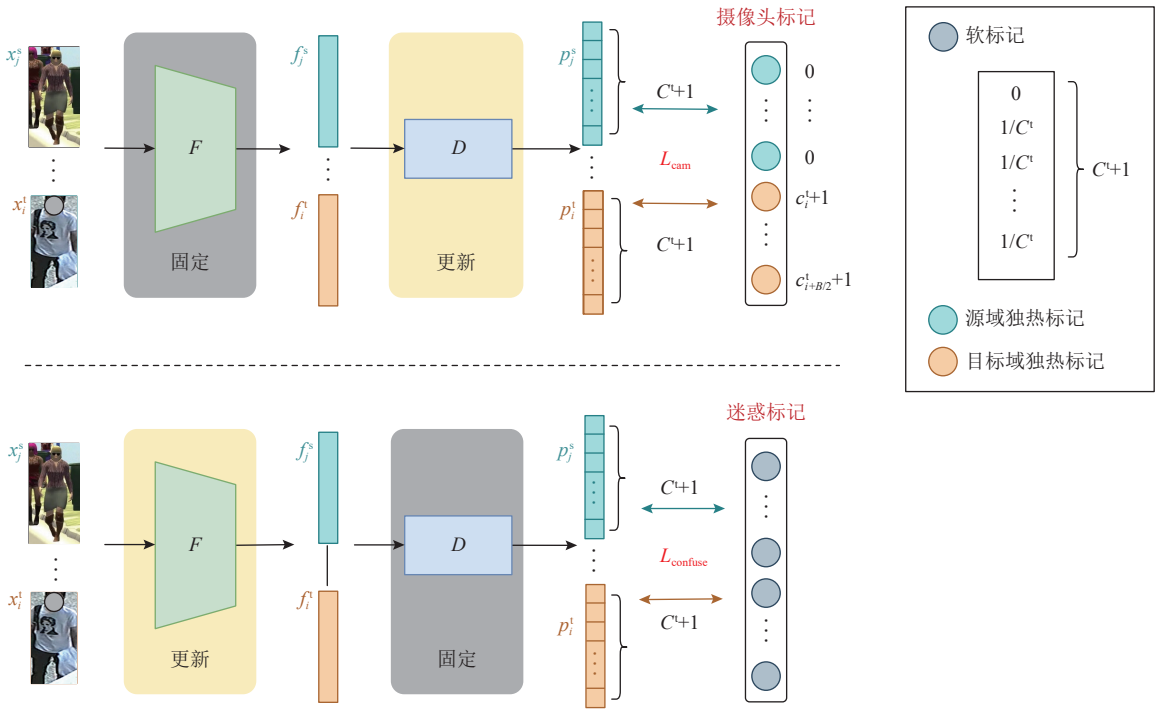


Fig. 4 Process of camera-aware based asymmetric adversarial learning

图4 基于摄像头感知的非对称对抗学习训练流程

征提取器参数，利用交叉熵损失只更新摄像头判别器，损失函数为：

$$\mathcal{L}_{\mathrm{cam}}=\frac{1}{N_{\mathrm{b}}}\sum_{i=1}^{N_{\mathrm{b}}}\mathcal{L}_{\mathrm{ce}}\left(D\left(F\left(\boldsymbol{x}_i^{\mathrm{b}}|\boldsymbol{\theta}_1\right)\right),c_i\right),\tag{6}$$

其中

$$c_i=\begin{cases}0,\ \boldsymbol{x}_i^{\mathrm{b}}\in X_{\mathrm{s}}.\\ c_i^{\mathrm{t}},\ \boldsymbol{x}_i^{\mathrm{b}}\in X_{\mathrm{t}}.\end{cases}$$

其次，为了将从属于不同摄像头的样本映射到共享子空间中，构建软标记，此时将摄像头判别器的参数固定，特征提取器的参数受身份损失与迷惑标记损失共同优化，迷惑损失为：

$$\mathcal{L}_{\mathrm{cfs}}=-\frac{1}{N_{\mathrm{b}}}\sum_{i=1}^{N_{\mathrm{b}}}\mathcal{L}_{\mathrm{ce}}\left(D\left(F\left(\boldsymbol{x}_i^{\mathrm{b}}|\boldsymbol{\theta}_1\right)\right),\boldsymbol{y}_{\mathrm{cfs}}\right),\tag{7}$$

其中$\boldsymbol{y}_{\mathrm{cfs}}$为迷惑标记，$\boldsymbol{y}_{\mathrm{cfs}}=\left[0,\frac{1}{C_{\mathrm{t}}},\frac{1}{C_{\mathrm{t}}},\cdots,\frac{1}{C_{\mathrm{t}}}\right]\in\mathbb{R}^{C_{\mathrm{t}}+1}$. 特征提取器和摄像头判别器以对抗的方式迭代学习，摄像头判别器要不断的通过优化$\mathcal{L}_{\mathrm{cam}}$，学习正确的摄像头标记信息来保证自己判别的准确性，而特征提取器需要生成出可以迷惑摄像头判别器的特征向量以此达到消除领域差异与领域内由摄像头带来的类内差异的问题，总优化损失为：

$$\mathcal{L}_{\mathrm{F}}=\lambda_{\mathrm{cam}}\mathcal{L}_{\mathrm{cfs}}+\mathcal{L}_{\mathrm{id}}^{\mathrm{mix}},\tag{8}$$

其中λ_{cam}为迷惑损失权重，$\mathcal{L}_{\mathrm{id}}^{\mathrm{mix}}$为由混合数据训练的通用行人重识别模型的身份损失. 通过控制合适的迷惑损失权重，可以平衡混合数据样本映射到共享子空时身份判别性的损失和在对抗中目标域数据分布发生被迫偏移的问题. 在对抗训练期间，特征提取器和摄像头判别器依次迭代学习，重复此操作直到收敛. 传统的双域对抗学习[8]只是这种基于摄像头感知的方案的一个特例，即当源域和目标域中只有 1 个摄像头类时.

3.4 辅助网络

已经有很多研究表明[42-43]，使用软标记作为分类交叉熵损失的推广是有效的. 同时受文献 [44] 的启发，我们认为在没有蒸馏的情况下，神经网络的训练会陷入部分特征的陷阱，因此需要引入知识蒸馏来缓解这个问题. 在实验中发现，对抗过程中，真实世界数据集的分布不可避免地会向虚拟数据集的数据分布偏移，这是我们不希望发生的. 为了避免这种偏移，本文将只由真实世界数据集训练的辅助网络作为教师网络，由辅助模型预测的身份类别概率作为软标记，代替目标域样本的独热标记，用于训练主模型中目标域数据身份分类分支，来缓解由虚拟数据集夹带的误导信息带来的不利影响.

如 3.1 节中介绍，辅助模型为由目标域数据训练的通用全监督行人重识别模型，然而在文献 [10] 中提到，虽然在教师网络中增加标记平滑的优化可以提高教师网络的模型性能，但是在用平滑过的教师网络进行知识蒸馏时，会降低学生网络的模型性能. 因为错误类别的预测概率正是我们在知识蒸馏时想要学习的潜在信息，而标记平滑则会抹去这些信息，所以为保证辅助模型预测的身份类别概率的可靠性，以及更好的进行知识蒸馏，辅助模型身份损失中的软分类损失由传统交叉熵损失代替进行更新优化，如式(9)所示：

$$\mathcal{L}_{\mathrm{id}}^{\mathrm{t}}=\mathcal{L}_{\mathrm{cls}}^{\mathrm{t}}+\mathcal{L}_{\mathrm{tri}}^{\mathrm{t}}.\tag{9}$$

此外，为了防止辅助模型预测样本身份类别错误导致主模型学习到错误信息. 本文构建了选择掩码矩阵$\boldsymbol{M}_{\mathrm{right}}\in\mathbb{R}^{1\times(N_{\mathrm{b}}/2)}$，在每个小批次中，只对在辅助模型中身份类别预测正确的样本计算对齐损失，选择掩码矩阵$\boldsymbol{M}_{\mathrm{right}}$和目标域对齐损失定义为：

$$\begin{aligned}\mathcal{L}_{\mathrm{alignment}}=-\frac{1}{m}\sum_{i=1}^{m}m_{\mathrm{r}}(I_{\mathrm{main}}(F\left(\boldsymbol{x}_i^{\mathrm{t}}|\boldsymbol{\theta}_1\right))\times\\ \log(I_{\mathrm{aux}}(F\left(\boldsymbol{x}_i^{\mathrm{t}}|\boldsymbol{\theta}_2\right))))\end{aligned}\tag{10}$$

其中

$$m_{\mathrm{r}}=\begin{cases}1,\ Index\left(\mathrm{argmax}I_{\mathrm{aux}}\left(F\left(\boldsymbol{x}_i^{\mathrm{t}}|\boldsymbol{\theta}_2\right)\right)\right)=i.\\ 0,\ Index\left(\mathrm{argmax}I_{\mathrm{aux}}\left(F\left(\boldsymbol{x}_i^{\mathrm{t}}|\boldsymbol{\theta}_2\right)\right)\right)\neq i.\end{cases}$$

辅助模型中身份分类器I_{aux}和主模型中身份分类器I_{main}均输出N_{t}维概率向量来预测目标域训练样本的身份. 对于一个样本$\boldsymbol{x}_i^{\mathrm{t}}$，当$I_{\mathrm{aux}}$输出的$N_{\mathrm{t}}$维向量中具有最大概率的索引值与它对应的身份标记的索引值i相同时，m_{r}的值为1.m为每个批次中选择掩码$\boldsymbol{M}_{\mathrm{right}}$中包含 1 的个数.

3 5 总训练流程

本文提出的基于知识迁移的虚实结合行人重识别方法框架包含由混合数据训练的主模型和由目标域数据单独训练的辅助模型. 在一个批次中首先由$\mathcal{L}_{\mathrm{cam}}$训练主模型中的摄像头判别器，随后总损失$\mathcal{L}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)$通过式(11)同时更新除摄像头判别器以外的主模型和辅助模型参数：

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)=\mathcal{L}_{\mathrm{id}}^{\mathrm{mix}}(\boldsymbol{\theta}_1)+\lambda_{\mathrm{cam}}\mathcal{L}_{\mathrm{cfs}}(\boldsymbol{\theta}_1)+\mathcal{L}_{\mathrm{id}}^{\mathrm{t}}(\boldsymbol{\theta}_2)+\\ \lambda_{\mathrm{aux}}\mathcal{L}_{\mathrm{alignment}}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2),\end{aligned}\tag{11}$$

其中λ_{cam}，λ_{aux}为权重参数. 主模型中特征提取器与摄像头判别器以对抗的方式迭代学习，将混合数据样

本映射到共享子空间,辅助模型通过简单的知识蒸馏去平衡对抗中带来的分布偏移. 主模型与辅助模型共同优化,同时建模域间、域内差异,使模型从更丰富多样的数据中学到更一般化的类内差异特征表示. 具体算法如算法 1 所示:

算法 1. 基于知识迁移的虚实结合行人重识别算法.

输入: 虚拟数据集 $X_s=\left\{\left(\boldsymbol{x}_i^s,\boldsymbol{y}_i^s\right)\right\}_{i=1}^{N_s}$;

真实数据集 $X_t=\left\{\left(\boldsymbol{x}_i^t,\boldsymbol{y}_i^t,c_i^t\right)\right\}_{i=1}^{N_t}$;

ImageNet 上预训练模型 $F(\cdot|\boldsymbol{\theta})$;

批样本数量 N_b, 身份采样参数 P,K, 权重参数 $\lambda_{cam},\lambda_{aux}$, 标记平滑参数 ε, 三元组阈值 a;

输出: 经过优化的特征提取器 $F(\cdot|\boldsymbol{\theta}_1)$.

① for n in [1, *num_epochs*] do

② repeat

③ B_s,B_t 分别为每个小批次从 X_s,X_t 中按照 $P\times K$ 采样 $N_b/2$ 个图像;

④ $B_{mix}=[B_s,B_t]$, 将 B_{mix} 经过数据增强后输入到 $F\left(\boldsymbol{x}_i^{mix}\middle|\boldsymbol{\theta}_1\right)$ 中, B_t 经过数据增强后输入到 $F\left(\boldsymbol{x}_i^t\middle|\boldsymbol{\theta}_2\right)$ 中;

⑤ 根据式(6)单独更新摄像头判别器参数, 其他参数固定不参与更新;

⑥ 根据式(11)更新除摄像头判别器以外的参数;

⑦ until B_t 被枚举完;

⑧ end for

4 实 验

4.1 数据集与评价指标

本文引入了虚拟数据集 RandPerson[45] 分别与 2 个广泛使用的真实域行人重识别数据集相结合进行训练. 并且在 2 个数据集上评估了所提出的方法, 2 个数据集分别为 CUHK03[46] 和 MSMT[34].

Randperson 由 19 个摄像头捕获的 8 000 个身份的 132 145 张图像组成. 因为此数据集仅作为辅助训练引入, 本文的目的并不在意模型在此数据集上的表现, 所以并未划分查询集与图库集, 此数据集也不会参加模型验证与测试.

CUHK03 包含 28 193 张图像, 由 2 个摄像头拍摄的 1 467 个身份组成. 本文遵循与 Market1501 类似的新协议, 将 CUHK03 数据集分为 767 个身份的训练集和 700 个身份的测试集. 从每个相机中, 选择一个图像作为每个身份标记的查询, 其余图像用于构建图库集. 其中 767 个身份的 7 368 张图像用于训练. 查询集和图库集分别具有其余 700 个身份的 1 400 个和 5 328 个图像.

MSMT17(下文简称为 MSMT)由 15 个摄像头捕获的 4 101 个身份的 126 441 张图像组成. 训练集包含 1 041 个身份的 32 621 张图像. 并将 3 060 个身份的其余图像分配给查询集(11 659 个图像)和图库集(82 161 个图像).

在训练过程中, 我们仅使用每个数据集训练集中的图像和相机标签, 不使用任何其他注释信息. 模型评价指标由累积匹配性能(cumulative match characteristic, CMC)rank-1 和平均精度均值(mean average precision, mAP)评估.

4.2 实验设置

本文在主干模型和辅助模型上均采用 ResNet-50[47] 作为主干网络去提取特征, 并使用在 ImageNet[48] 上预先训练的参数对其进行初始化. 在训练过程中, 输入图像被统一调整为 256×128, 并采取了行人重识别领域大多数工作采取的传统图像增强方式, 是通过随机翻转、随机填充和随机擦除[30] 进行的.

对于训练中的每个小批次中, 每个身份的图像都遵循 $P\times K$ 采样. 为了确保与其他方法更公平的比较, 本文采取了同其他行人重识别方法[6,16,26] 的实验设置. 本文设定小批量大小为 128, 即对于混合数据来说, 采用均衡采样策略, 分别在虚拟数据集和真实世界数据集使用 $P=16$ 个随机选择的身份, 每个身份选择 $K=4$ 张图像进行采样, 主模型输入混合数据, 辅助模型输入与混合数据中相同的真实数据样本.

采用 Adam 方法对模型进行优化. 训练中采取预热策略来引导网络以获得更好的性能, 如图所示, 本文将初始学习率设置为 3.5×10^{-5}. 前 10 个 epoch 线性地将学习率从 3.5×10^{-5} 增加到 3.5×10^{-4}, 在第 40 个轮次和第 70 个轮次学习率分别衰减到 3.5×10^{-5} 和 3.5×10^{-6}, 总共有 120 个训练轮次. 三元组损失的间距阈值 a 设为 0.3. 本文在主干网络中引入了标签平滑技术, 并参考了基于标签平滑的知识蒸馏方法[42] 的实验设置, 将标签平滑参数 ε 设定为 0.1.

4.3 对比实验结果

本节中将本文提出的基于对抗迁移学习的行人重识别方法, 与近几年行人重识别领域提出的主流全监督方法在 CUHK03 和 MSMT17 这 2 个数据集上进行了比较, 这里需要注意的是, 本文没有做任何的

后处理，例如，重新排名[27-28]或多查询融合[1]，因此所有的比较是公平的.

在 CUHK03 数据集上的验证：通常有 2 种拆分方法可以将 CUHK03 数据集划分训练集和测试集. 本文采取了 767/700 的拆分设置（训练集包含 767 个身份标记、测试集包含 700 个身份标记）而非 1 367/100（训练集包含 1 367 个身份标记、测试集包含 100 个身份标记）的设置. 因为前者比后者具有更少的训练图像和更多的测试图像，设置上更难，具有更大的挑战性也更接近现实的行人重识别任务. 同时在这种设置下也会给模型带来更大的困难. 在这种情况下，引入无需手工标注的虚拟数据集可以很好的帮助模型进行训练. 由表 1 可以看出，本文所提出的方法在 CUHK03 数据集上无论采取哪种评价指标，与目前主流的方法对比均可以取得卓越性能.

Table 1　Compared with SOTA Methods on CUHK03

表 1　CUHK03 数据集上与主流方法的对比

方法	主干网络	尺寸	准确率/%	
			mAP	rank-1
PCB[4]	ResNet50	256×128	53.2	59.7
FD[49]	ResNet50	256×128	65.2	67.7
Pyramid[16]	ResNet101	384×128	76.9	78.9
MGN[6]	ResNet50	384×128	67.4	68.0
BAT-net[17]	GoogLeNet	256×128	76.1	78.6
DANN[8]*	ResNet50	256×128	75.5	77.6
AutoLoss-GMS-A[50]	ResNet50	256×128	68.3	70.4
本文	ResNet50	256×128	77.8	81.1

注：*表示在本文相同设置下重新实现.

Pyramid 方法它连接了 21 个不同尺度的局部特征，并且使用了更多参数的 ResNet101 作为主干网络.MGN 方法受益于更庞大且复杂的网络，它具有 8 个分支，生成 8 个特征向量，共需要 11 个损失函数进行监督约束. 其模型大小（参数数量）更是本文方法的 3 倍. 然而在 CUHK03 数据集上进行对比时，本文方法在 2 个评价指标上均高于以上 2 个方法. 这证明了在面对更困难、更具有挑战的行人重识别任务时引入虚拟数据集进行辅助训练的重要性，可以在更小的开销下提升更多的模型性能. 在与本文具有相同的数据集和采样策略的设置下重新实现了 DANN 方法. 重现的模型在 rank-1 上比 CAAL 方法降低约 3.4%，在 mAP 上降低约 2.3%，这也证明了本文方法在消除领域差异上的优越性.

在 MSMT 数据集上的验证说明了 MSMT 数据集是 3 个数据集中拥有图像数量最多、行人身份数目最多、摄像头数量最多的数据集. 该数据集具有更复杂的场景和背景，并且在采集图片时涵盖了多个天气，多个时段具有更复杂的光照变化. 因此 MSMT 数据集是更接近真实场景的大型数据集. 然而如表 2 所示，在面对如此庞大且复杂的数据集时，本文方法比对比方法的 rank-1 值高出至少 1.7% 的和比对比方法的 mAP 值高出至少 1.3%.

Table 2　Compared with SOTA Methods on MSMT Dataset

表 2　MSMT 数据集上对比主流方法

方法	主干网络	尺寸	准确率/%	
			mAP	rank-1
PCB[4]	ResNet50	256×128	40.4	68.2
BAT-net[17]	GoogLeNet	256×128	56.8	79.5
CE-FAT[51]	ResNet50	256×128	39.2	69.4
CBN[26]	ResNet50	256×128	42.9	72.8
DG-Net[35]	ResNet50	256×128	52.3	77.2
DANN[8]*	ResNet50	256×128	57.6	79.9
CDNet[7]	CDNet	256×128	54.7	78.9
AutoLoss-GMS-A[50]	ResNet50	256×128	55.1	79.5
本文	ResNet50	256×128	58.8	82.1

注：*表示在本文相同设置下重新实现.

CBN 方法中也利用到了摄像头标记信息，但是仍低于本文方法的 mAP 值 15.9% 和 rank-1 值 9.3%，证明本文方法可以更好地利用摄像头标记信息.DG-Net 生成了 10 倍的数据增广，在 mAP 和 rank-1 这 2 个指标上分别低了 6.5% 和 4.9%，证明了引入虚拟数据集进行辅助训练的有效性. 与 CUHK03 数据集相同，在 MSMT 数据集上同样复现了与本文方法框架具有相同的数据集和采样策略的设置的 DANN 方法. 在此设置下复现的 DANN 方法依然在 mAP 上略高于表现最好的 BAT-net 方法，这也证明了即使是面对数据量更多、行人身份更多样、场景更复杂的大型数据集时，引入虚拟数据集进行辅助训练仍然对模型表现具有卓越的提升. 同时，将本文方法与复现的 DANN 方法对比时，mAP 值提高了 1.2%，rank-1 值提高了 2.2%，证明了本文消除领域差异的方法更适用于行人重识别任务.

综上所述，本文方法在 2 个数据集上与当前主流行人重识别方法对比，无论使用哪种评估指标均表现优异. 通过对 2 个数据集的比较实验，我们也可以发现本文方法在更具有挑战性的新协议 CUHK03 数据集和数据量更大更复杂的 MSMT 数据集上进行验

证时,始终优于所有其他算法.因此,我们可以得出本文方法更擅长处理复杂且具有挑战性的问题,即更接近真实场景的情况.

4.4 消融实验分析

本节中我们在 CUHK03 和 MSMT 数据集上进行了消融实验,以充分分析和验证每个组件的有效性.设置 ResNet-50 为基准网络,其中“RP+CUHK03/MSMT”表示引入 RandPerson 数据集进行辅助训练,“w/o 辅助网络”表示仅使用我们提出的 CAAL 进行训练,“w/o CAAL”表示仅使用我们提出的辅助网络进行训练.

由于虚拟数据集 RandPerson 具有 132 145 张训练图像,是 CUHK03 数据集的 20 倍,即使是大型数据集 MSMT 也有将近 4 倍的训练数据.如果单纯将虚拟数据集和真实数据集(CUHK03 和 MSMT 中任意一个数据集)混合起来,每个小批次时随机从混合数据集中采样,不仅会造成训练时间的增长(训练数据的大幅增加),并且在实验中我们发现,直接混合训练不能提升模型效果,反而在 2 个真实数据集上有不同程度的下降,具体数据见表 3 和表 4.通过分析,造成这种现象的原因有 2 点:一是因为虚拟数据集与真实数据集具有较大的领域差异,强行混合训练只会令模型迷惑;二是虚拟数据集具有更多的数据量,从而使模型在学习的时候更偏向于虚拟数据集的数据分布.

Table 3 Ablation Experiment Results on CUHK03 Dataset

表 3 CUHK03 数据集上消融实验结果

方法	训练集	准确率/%	
		mAP	rank-1
基准模型	CUHK03	63.6	65.7
基准模型	RP+ CUHK03	63.8	65.6
基准模型*	RP+ CUHK03	72.7	75.1
本文(w/o 辅助网络)	RP+ CUHK03	77.1	80.5
本文(w/o CAAL)	RP+ CUHK03	75.9	78.2
本文(完整方法)	RP+ CUHK03	77.8	81.1

注:*表示采用均衡采样策略.

Table 4 Ablation Experiment Results on MSMT Dataset

表 4 MSMT 数据集上消融实验结果

方法	训练集	准确率/%	
		mAP	rank-1
基准模型	MSMT	49.8	75.1
基准模型	RP+ MSMT	47.8	71.9
基准模型*	RP+ MSMT	56.3	78.9
本文(w/o 辅助网络)	RP+ MSMT	57.3	81.0
本文(w/o CAAL)	RP+ MSMT	57.8	80.3
本文(完整方法)	RP+ MSMT	58.8	82.1

注:*表示采用均衡采样策略.

为了缓减上述第 2 个问题,我们采用均衡采样策略生成每一个批训练数据,即一个批中的数据有一半真实数据和一半虚拟数据.然而传统的方法,直接随机从每个数据集中进行采样训练.从表 3 和表 4 中可以看出,均衡采样策略能够显著提升模型的性能,在我们的方法中,我们也使用该采样策略.

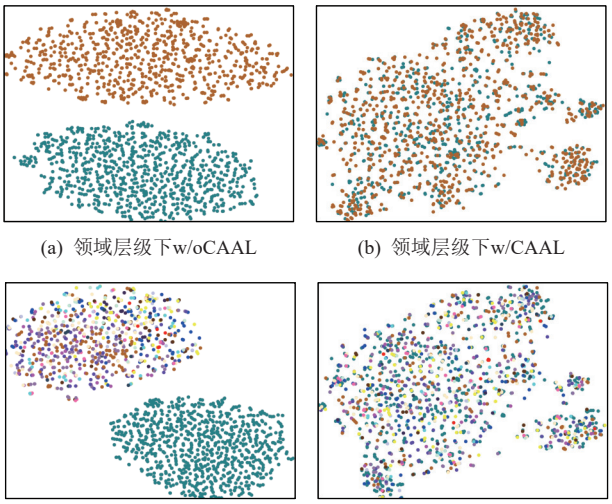
为了证明基于摄像头感知的非对称领域对抗学习模块的有效性,表 3、表 4 展示了不使用该模块和使用该模块的实验结果,可以看出当使用 CAAL 模块时在 CUHK03 数据集上的 mAP 值提升了 4.4%、rank-1 值提升了 5.4%,在 MSMT 数据集上 mAP 值和 rank-1 值分别提升 1.0% 和 2.1%.考虑到 CUHK03 是更复杂,从这也可以看出 CAAL 模块在更困难(更逼近真实场景)的任务上作用更明显.

为了证明辅助网络的有效性,表 3、表 4 同样展示了不使用该模块和使用该模块的实验结果,当仅使用辅助网络时,2 个数据集在 mAP 和 rank-1 指标上均有提升,证明引入知识蒸馏的确可以在一定程度上缓解虚拟数据集夹带的误导信息带来的不利影响.当我们在加入 CAAL 的方法上加入辅助网络时,实验结果表明在 2 个数据集上的性能均有显著的提升,证明了模型效果的提升来自于 CAAL 模块和辅助网络的共同作用.

为了进一步证明 CAAL 模块可以同时降低领域差异和多摄像头间的类内差异,本节采用的 t-SNE[52] 可视化技术分别绘制了在引入 RandPerson 数据集辅助 MSMT 数据集时领域层级和摄像头层级的数据分布示意图.其中,图 5(a)(b)分别是领域层级下不使用 CAAL 和使用 CAAL 时的数据分布,图 5(c)(d)分别是摄像头层级下不使用 CAAL 和使用 CAAL 时的数据分布.

图 5(a)(b)中,所有的点分别代表了数据集 RandPerson 和 MSMT 的样本.可以看出在不使用 CAAL 时这 2 个数据集具有较大的领域差异,而在使用 CAAL 后 2 个数据集的样本混合在了一起,即领域差异被消除了.

而在图 5(c)(d)中,将虚拟数据集看作从属于一个摄像头下,所有的点分别代表了 RandPerson 和 MSMT 15 个摄像头下每个摄像头的样本.可以看出在没有使用 CAAL 时,MSMT 中的摄像头样本更倾向于聚集在左上部分,RandPerson 的摄像头样本更倾



(a) 领域层级下w/oCAAL (b) 领域层级下w/CAAL

(c) 摄像头层级下w/oCAAL (d) 摄像头层级下w/CAAL

Fig. 5 Visualization of data distribution via t-SNE

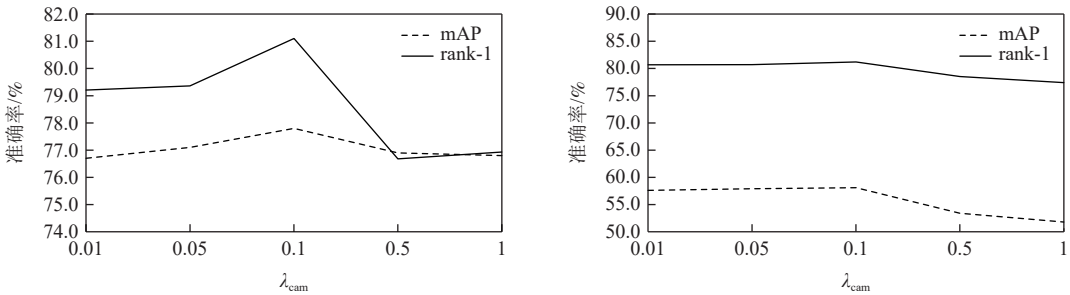
图 5 t-SNE 可视化数据分布

向于聚集在右下部分，这体现了不同摄像头通常具有不同的数据分布，当同一身份样本出现在不同摄像头下就会出现类内差异. 而使用了 CAAL 后不仅虚拟数据集与真实世界数据集的领域差异消失了，原本由于不同摄像头都具有特有的数据分布而聚集在一起的样本已经均匀分散开了. 这表明了本文所提出的 CAAL 模块消除领域差异的同时也弱化了多摄像头间的类内差异.

4.5 参数实验分析

图 6 展示了本文在 2 个数据集上针对损失权重 λ_{cam} 进行的参数实验，可以看出当损失权重 $\lambda_{\text{cam}} \geqslant 0.5$ 的时候 2 个数据集都出现了不同程度的下降，这也证实了我们之前的想法，即使我们的目标是让虚拟数据集的分布尽可能的靠近真实世界数据集分布，但是在对抗中真实世界数据集分布仍然会向虚拟数据集偏移，实验证明，当权重参数 $\lambda_{\text{cam}} = 0.1$ 时，可以平衡混合数据样本映射到共享子空间时身份判别性的损失和在对抗中目标域数据分布发生被迫偏移的问题.

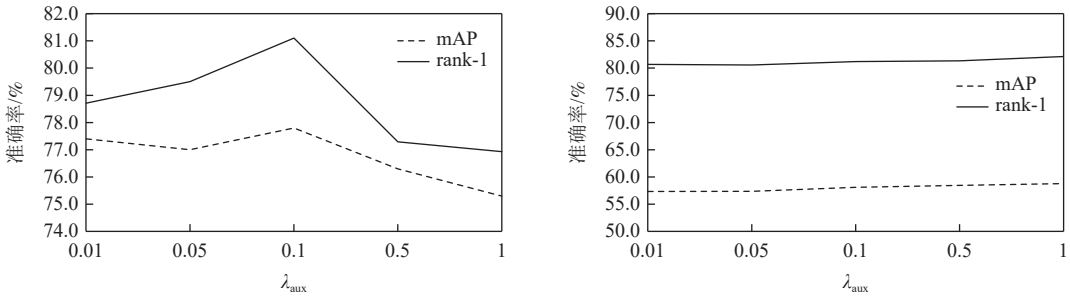
本文也对辅助网络中的权重参数 λ_{aux} 进行了参数实验，实验结果见图 7. 我们发现对虚拟数据集依赖性弱的数据集（例如，MSMT 数据集在直接混合虚拟数据集训练时，相比于仅依靠自身数据训练时在 2 个指标都有不同程度的下降），此时由虚拟数据集夹带的误导信息带来的不利影响越多，辅助网络所带来的知识蒸馏效果越好，所带来的提升也越高，所以将权重参数 λ_{aux} 设置为 1.0. 而数据集是较为困难的任务时（例如，CUHK03 数据集在直接混合虚拟数据集训练时与仅依靠自身数据训练时相比 mAP 指标略有上升），此时辅助网络作为教师网络的可信度下降，



(a) CUHK03 (b) MSMT

Fig. 6 Analysis of λ_{cam} on CUHK03 and MSMT

图 6 CUHK03 和 MSMT 上 λ_{cam} 的分析



(a) CUHK03 (b) MSMT

Fig. 7 Analysis of λ_{aux} on CUHK03 and MSMT

图 7 CUHK03 和 MSMT 上 λ_{aux} 的分析

当模型过于依赖辅助网络时反而会对任务造成负迁移效果. 此时,实验中我们将权重参数λ_{aux}设置为0.1,这样模型在应对数据量较少且更接近现实场景的行人重识别任务时就不会过度依赖辅助网络,但是同时又可以从辅助网络蒸馏的知识中学到原有的数据分布,以此缓解虚拟数据集带来的知识偏移.

5 结 论

本文根据现有全监督行人重识别任务受限于数据匮乏的问题,提出引用虚拟数据集进行辅助训练. 同时提出了一种基于知识迁移的虚实结合行人重识别方法框架,采用基于摄像头感知的非对称领域对抗学习方法同时缓解域间、域内差异,以解决虚拟数据集与真实世界数据集领域差异过大和由不同摄像头带来的类内差异问题. 实验证明了该方法在逼近真实场景的行人重识别任务上的有效性.

作者贡献声明: 祁磊提出了算法思路和实验方案;任子豪负责论文修订、补充和实验分析;刘俊汐执行实验和撰写论文;耿新提出指导意见并修改论文.

参 考 文 献

[1] Zheng Liang, Shen Liyue, Tian Lu, et al. Scalable person re-identification: a benchmark[C]//Proc of the 15th IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2015: 1116−1124

[2] Ristani E, Solera F, Zou R, et al. Performance measures and a data set for multi-target, multi-camera tracking[C]//Proc of the 14th European Conf on Computer Vision. Berlin: Springer, 2016: 17−35

[3] Zhang Xuan, Luo Hao, Fan Xing, et al. Alignedreid: surpassing human-level performance in person re-identification[J]. arXiv preprint, arXiv: 1711.08184, 2017

[4] Sun Yifan, Zheng Liang, Yang Yi, et al. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)[C]// Proc of the 15th European Conf on Computer Vision. Berlin: Springer, 2018: 480−496

[5] Luo Hao, Jiang Wei, Gu Youzhi, et al. A strong baseline and batch normalization neck for deep person re-identification[J]. IEEE Transactions on Multimedia, 2019, 22(10): 2597−2609

[6] Wang Guanshuo, Yuan Yufeng, Chen Xiong, et al. Learning discriminative features with multiple granularities for person re-identification[C]//Proc of the 26th ACM Int Conf on Multimedia. New York: ACM, 2018: 274−282

[7] Li Hanjun, Wu Gaojie, Zheng Weishi. Combined depth space based architecture search for person re-identification[C]//Proc of the 34th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 6729−6738

[8] Ganin Y, Ustinova E, Ajakan H, et al. Domain-adversarial training of neural networks[J]. The Journal of Machine Learning Research, 2016, 17(1): 2096−2030

[9] Das A, Chakraborty A, Roy-Chowdhury A K. Consistent re-identification in a camera network[C]//Proc of the 13th European Conf on Computer Vision. Berlin: Springer, 2014: 330−345

[10] Liao Shengcai, Hu Yang, Zhu Xiangyu, et al. Person re-identification by local maximal occurrence representation and metric learning[C]// Proc of the 28th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 2197−2206

[11] Li Zhen, Chang Shiyu, Liang Feng, et al. Learning locally-adaptive decision functions for person verification[C]//Procs of the 26th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2013: 3610−3617

[12] Yang Yang, Yang Jimei, Yan Junjie, et al. Salient color names for person re-identification[C]//Proc of the 13th European Conf on Computer Vision. Berlin: Springer, 2014: 536−551

[13] Chen Liwen, Ye Feng, Huang Tianqiang, et al. An unsupervised person re-identification method based on intra-/inter-camera merger[J]. Journal of Computer Research and Development. 2023, 60(2): 415−425 (in Chinese)
(陈利文,叶锋,黄添强,等. 基于摄像头域内域间合并的无监督行人重识别方法[J]. 计算机研究与发展,2023,60(2):415−425)

[14] Varior R R, Shuai Bing, Lu Jiwen, et al. A siamese long short-term memory architecture for human re-identification[C]//Proc of the 14th European Conf on Computer Vision. Berlin: Springer, 2016: 135−153

[15] Wei Longhui, Zhang Shiliang, Yao Hantao, et al. Glad: Global-local-alignment descriptor for pedestrian retrieval[C]//Proc of the 25th ACM Int Conf on Multimedia. New York: ACM, 2017: 420−428

[16] Zheng Feng, Deng Cheng, Sun Xing, et al. Pyramidal person re-identification via multi-loss dynamic training[C]//Proc of the 32nd IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 8514−8522

[17] Fang Pengfei, Zhou Jieming, Roy S K, et al. Bilinear attention networks for person retrieval[C]//Proc of the 32nd IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 8030−8039

[18] Shi Yuexiang, Zhou Yue. Person re-identification based on stepped feature space segmentation and local attention mechanism[J]. Journal of Electronics & Information Technology, 2022, 44(1): 195−202 (in Chinese)
(石跃祥,周玥. 基于阶梯型特征空间分割与局部注意力机制的行人重识别[J]. 电子与信息学报,2022,44(1):195−202)

[19] Li Wei, Zhu Xiatian, Gong Shaogang. Harmonious attention network for person re-identification[C]//Proc of the 31st IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 2285−2294

[20] Zhao Liming, Li Xi, Zhuang Yueting, et al. Deeply-learned part-aligned representations for person re-identification[C]//Proc of the 16th IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2017:

3219-3228

[21] Kim W, Goyal B, Chawla K, et al. Attention-based ensemble for deep metric learning[C]// Proc of the 15th European Conf on Computer Vision. Berlin: Springer, 2018: 736−751

[22] Li Dangwei, Chen Xiaotang, Zhang Zhang, et al. Learning deep context-aware features over body and latent parts for person re-identification[C]// Proc of the 30th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 384−393

[23] Kumar V, Namboodiri A, Paluri M, et al. Pose-aware person recognition[C]//Proc of the 30th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 6223-6232

[24] Su Chi, Li Jianning, Zhang Shiliang, et al. Pose-driven deep convolutional model for person re-identification[C]//Proc of the 16th IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2017: 3960−3969

[25] Zhong Zhun, Zheng Liang, Zheng Zhedong, et al. Camera style adaptation for person re-identification[C]//Proc of the 31st IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 5157−5166

[26] Zhuang Zijie, Wei Longhui, Xie Lingxi, et al. Rethinking the distribution gap of person re-identification with camera-based batch normalization[C]// Proc of the 16th European Conf on Computer Vision. Berlin: Springer, 2020: 140−157

[27] Zhong Zhun, Zheng Liang, Cao Donglin, et al. Re-ranking person re-identification with k-reciprocal encoding[C]//Proc of the 30th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 1318−1327

[28] Bai Song, Bai Xiang, Tian Qi. Scalable person re-identification on supervised smoothed manifold[C]//Proc of the 30th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 2530−2539

[29] Schumann A, Stiefelhagen R. Person re-identification by deep learning attribute-complementary information[C]//Proc of the 30th IEEE Conf on Computer Vision and Pattern Recognition Workshops. Piscataway, NJ: IEEE, 2017: 20−28

[30] Zhong Zhun, Zheng Liang, Kang Guoliang, et al. Random erasing data augmentation[C]//Proc of the 34th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2020, 34(7): 13001−13008

[31] DeVries T, Taylor G W. Improved regularization of convolutional neural networks with cutout[J]. arXiv preprint, arXiv: 1708.04552, 2017

[32] Liu Jinxian, Ni Bingbing, Yan Yichao, et al. Pose transferrable person re-identification[C]//Proc of the 31st IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 4099−4108

[33] Qian Xuelin, Fu Yanwei, Xiang Tao, et al. Pose-normalized image generation for person re-identification[C]//Proc of the 15th European Conf on Computer Vision. Berlin: Springer, 2018: 650−667

[34] Wei Longhui, Zhang Shiliang, Gao Wen, et al. Person transfer gan to bridge domain gap for person re-identification[C]//Proc of the 31st IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 79-88

[35] Zheng Zhedong, Yang Xiaodong, Yu Zhiding, et al. Joint discriminative and generative learning for person re-identification[C]//Proc of the 32nd IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 2138−2147

[36] Yu Chaohui, Wang Jindong, Chen Yiqiang, et al. Transfer learning with dynamic adversarial adaptation network[C]//Proc of the 19th IEEE Int Conf on Data Mining. Piscataway, NJ: IEEE, 2019: 778−786

[37] Qi Lei, Wang Lei, Huo Jing, et al. A novel unsupervised camera-aware domain adaptation framework for person re-identification[C]// Proc of the 32nd IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 8080−8089

[38] Dai Chenchao, Wang Hongyuan, Ni Tongguang, et al. Person re-identification based on deep convolutional generative adversarial network and expanded neighbor reranking[J]. Journal of Computer Research and Development, 2019, 56(8): 1632−1641 (in Chinese)
(戴臣超，王洪元，倪彤光，等. 基于深度卷积生成对抗网络和拓展近邻重排序的行人重识别[J]. 计算机研究与发展，2019，56(8)：1632−1641)

[39] Wang Wenhao, Liao Shengcai, Zhao Fang, et al. Domainmix: Learning generalizable person re-identification without human annotations[J]. arXiv preprint, arXiv: 2011.11953, 2020

[40] Zhang Ying, Xiang Tao, Hospedales T M, et al. Deep mutual learning[C]//Proc of the 31st IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 4320−4328

[41] Hermans A, Beyer L, Leibe B. In defense of the triplet loss for person re-identification[J]. arXiv preprint, arXiv: 1703.07737, 2017

[42] Müller R, Kornblith S, Hinton G E. When does label smoothing help[J]. Advances in Neural Information Processing Systems, 2019, 32: 4694−4703

[43] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network[J]. arXiv preprint, arXiv: 1503.02531, 2015

[44] Allen-Zhu Z, Li Yuanzhi. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning[J]. arXiv preprint, arXiv: 2012.09816, 2020

[45] Wang Yanan, Liao Shengcai, Shao Ling. Surpassing real-world source training data: Random 3D characters for generalizable person re-identification[C]// Proc of the 28th ACM Int Conf on Multimedia. New York: ACM, 2020: 3422−3430

[46] Li Wei, Zhao Rui, Xiao Tong, et al. Deepreid: Deep filter pairing neural network for person re-identification[C]//Proc of the 27th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2014: 152−159

[47] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Deep residual learning for image recognition[C]//Proc of the 29th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 770−778

[48] Deng Jia, Dong Wei, Socher R, et al. Imagenet: A large-scale hierarchical image database[C]//Proc of the 22nd IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2009: 248−255

[49] Ren Pengyuan, Li Jianmin. Factorized distillation: Training holistic

person re-identification model by distilling an ensemble of partial ReID models[J]. arXiv preprint, arXiv: 1811.08073, 2018

[50] Gu Hongyang, Li Jianmin, Fu Guangyuan, et al. Autoloss-gms: Searching generalized margin-based softmax loss function for person re-identification[C]//Proc of the 35th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 4744−4753

[51] Yuan Ye, Chen Wuyang, Yang Yang, et al. In defense of the triplet loss again: Learning robust person re-identification with fast approximated triplet loss and label distillation[C]//Proc of the 33rd IEEE Conf on Computer Vision and Pattern Recognition Workshops. Piscataway, NJ: IEEE, 2020: 354−355

[52] Van der Maaten L, Hinton G. Visualizing data using t-SNE[J]. Journal of machine learning research, 2008, 9(11): 2579−2605

Qi Lei, born in 1989. PhD, assistant researcher. Member of CCF. His main research interests include pattern recognition and computer vision.

祁　磊, 1989 年生. 博士, 助理研究员. CCF 会员. 主要研究方向为模式识别、计算机视觉.

Ren Zihao, born in 2000. Master candidate. His main research interests include pattern recognition and computer vision.

任子豪, 2000 年生. 硕士研究生. 主要研究方向为模式识别、计算机视觉.

Liu Junxi, born in 1998. Master candidate. Her main research interests include pattern recognition and computer vision.

刘俊汐, 1998 年生. 硕士研究生. 主要研究方向为模式识别和计算机视觉.

Geng Xin, born in 1978. PhD, professor. Member of CCF. His main research interests include machine learning and computer vision.

耿　新, 1978 年生. 博士, 教授. CCF 会员. 主要研究方向为机器学习、计算机视觉.