基于自注意力模态融合网络的跨模态行人再识别方法研究

杜 鹏 [1] 宋永红 [1, 2] 张鑫瑶 [1]

摘 要 行人再识别是实现多目标跨摄像头跟踪的核心技术, 该技术能够广泛应用于安防、智能视频监控、刑事侦查等领域. 一般的行人再识别问题面临的挑战包括摄像机的低分辨率、行人姿态变化、光照变化、行人检测误差、遮挡等. 跨模态行人再识别相比于一般的行人再识别问题增加了相同行人不同模态的变化. 针对跨模态行人再识别中存在的模态变化问题, 本文提出了一种自注意力模态融合网络. 首先是利用 CycleGAN 生成跨模态图像. 在得到了跨模态图像后利用跨模态学习网络同时学习两种模态图像特征, 对于原始数据集中的图像利用 SoftMax 损失进行有监督的训练, 对生成的跨模态图像利用 LSR (Label smooth regularization) 损失进行有监督的训练. 之后, 使用自注意力模块将原始图像和 CycleGAN 生成的图像进行区分, 自动地对跨模态学习网络的特征在通道层面进行筛选. 最后利用模态融合模块将两种筛选后的特征进行融合. 通过在跨模态数据集 SYSU-MM01 上的实验证明了本文提出的方法和跨模态行人再识别其他方法相比有一定程度的性能提升.

关键词 跨模态行人再识别, 自注意力, 跨模态融合, CycleGAN

引用格式 杜鹏, 宋永红, 张鑫瑶. 基于自注意力模态融合网络的跨模态行人再识别方法研究. 自动化学报, 2022, **48**(6): 1457−1468

DOI 10.16383/j.aas.c190340

Self-attention Cross-modality Fusion Network for Cross-modality Person Re-identification

DU Peng[1] SONG Yong-Hong[1, 2] ZHANG Xin-Yao[1]

Abstract Person re-identification is the core technology to achieve multi-target multi-camera tracking. It can be widely used in many areas such as security, intelligent video surveillance, and criminal investigation. Person re-identification is a challenging task due to the low resolution of camera, human pose variations, illumination variations, pedestrian detector errors and occlusion. Compared with the general person re-identification, the cross-modality person re-identification has the variations of different modalities of the same person. In order to solve the cross-modality problem in cross-modality person re-identification, we propose the self-attention cross-modality fusion network. First, CycleGAN is used to generate cross-modality images. After obtaining the cross-modality images, we use the cross-modality learning network to learn the two modalities features simultaneously. SoftMax loss is used to train original images and label smooth regularization (LSR) loss is used to train generated images. Then, we use self-attention module to distinguish between original images and the generated image, and automatically select the useful features between channels. Finally, modality fusion module is used to fuse these selected features from two modalities images. Comparing with state-of-the-art methods on a large scale cross-modality dataset SYSU-MM01 further demonstrate the effectiveness of the proposed self-attention cross-modality fusion network.

Key words Cross-modality person re-identification, self-attention, cross-modality fusion, CycleGAN

Citation Du Peng, Song Yong-Hong, Zhang Xin-Yao. Self-attention cross-modality fusion network for cross-modality person re-identification. *Acta Automatica Sinica*, 2022, **48**(6): 1457−1468

近年来, 伴随着视频采集技术的大力发展, 大量的监控摄像头部署在商场、公园、学校等公共场所. 监控摄像的出现给人们带来了极大的便利, 其中最直接的一个好处就是可以帮助公安等执法部门解决盗窃、抢劫等重大刑事案件. 但是正是由于监控摄像头布置的区域十分广阔, 基本在大中小城市中都遍地布满了监控摄像头, 当一个目标人物在一个城市的监控摄像网络中移动时, 往往会导致公安等相关部门人员在一定时间内在整个网络中对监控视频进行查看, 这对公安等相关部门进行区域的管理以及视频的查看带来了较大的不便. 因此, 需要一种方便、快捷的方式来代替人工对监控视频中行人进行搜寻. 为了实现对监控视频中的行人进行搜

收稿日期 2019-05-07 录用日期 2019-10-11
Manuscript received May 7, 2019; accepted October 11, 2019
国家重点研究发展计划 (2017YFB1301101), 陕西省自然科学基础研究计划 (2018JM6104) 资助
Supported by National Key Research and Development Program of China (2017YFB1301101) and Natural Science Basic Research Program of Shaanxi Province (2018JM6104)
本文责任编委 白翔
Recommended by Associate Editor BAI Xiang
1. 西安交通大学软件学院 西安 710049 2. 西安交通大学人工智能学院 西安 710049
1. School of Software Engineering, Xi′an Jiaotong University, Xi′an 710049 2. College of Artificial Intelligence, Xi′an Jiaotong University, Xi′an 710049

寻这个目标, 其本质就是要实现多目标跨摄像头追踪, 而行人再识别技术[1−2] 是多目标跨摄像头追踪问题的核心与关键. 行人再识别和多目标跨摄像头追踪的关系如图 1 所示. 实际场景中, 摄像头拍摄到的是包含众多行人与复杂背景的图像, 这个时候可以利用行人检测技术从拍摄到的复杂全景图像中得到行人包围框, 之后对于行人包围框集合利用行人再识别技术进行搜寻.

图 1 行人再识别和多目标跨摄像头跟踪关系示意

Fig. 1 The relationship between person re-identification and multi-target cross-camera tracking

除此之外, 犯罪分子通常会在夜间行动, 这时仅仅靠 RGB 相机去采集图像不能很好地解决这种夜间出现的行人匹配问题. 为了对夜晚出现的行人也能进行匹配, 除了 RGB 相机外, 有些地方可能会布控红外 (Infrared, IR) 相机, 这样, 在夜间或者是光线较暗处也可以采集到行人的红外图, 弥补了在夜晚传统的 RGB 相机采集失效的问题. 在这种情况下, RGB 图和 IR 图之间的跨模态匹配 (跨模态行人再识别) 具有很重要的现实意义. 跨模态匹配的重点是寻找不同模态间的相似性[3−4], 从而跨越模态对行人再识别的限制.

跨模态行人再识别相对于传统的行人再识别, 除了面临行人之间姿态变化、视角变化等问题外, 数据之间还存在跨模态的难点. 图 2 为跨模态行人再识别数据集中的行人数据. 图中第 1 行为在白天通过 RGB 相机在室内采集到的 RGB 图像; 第 2 行为在夜晚通过红外相机在室内采集到的 IR 图像; 第 3 行为白天在室外采集到的 RGB图像; 第 4 行为夜晚在室外采集到的 IR 图像. 每一列的 4 张图片属于同一个人, 不同列的图片属于不同的人. 与传统的 RGB-RGB 图像之间的匹配不同, 跨模态数据集上所关注的是 IR 图像和 RGB 图像之间的匹配, 这种跨模态匹配为行人再识别增加了不少难度, 如图 2 中第 3 列和第 4 列的两个行人, 通过 RGB 图可以很好地进行区分, 但通过 IR 图和 RGB 图匹配, 难度有一定程度的提升.

针对上述这些问题, 本文主要创新点如下:

1) 提出一种自注意力模态融合网络以解决跨模态行人再识别中存在的模态变化问题;

2) 提出使用 CycleGAN 对图像进行模态间的转换, 从而解决学习时需要对应的样本对问题;

3) 提出使用自注意力机制进行不同模态之间的特征筛选, 从而有效地对原始图像和使用 CycleGAN 生成的图像进行区分.

图 2 跨模态行人再识别数据

Fig. 2 Data of cross-modality person re-identification

1 相关方法概述

1.1 RGB-RGB 匹配的行人再识别方法

近年来, 随着模式识别以及深度学习的发展, 研究人员针对行人再识别方法做了大量的实验与研究工作. 前期针对行人再识别的方法主要集中于利用传统的模式识别方法, 例如设计行人特征来表示行人, 或者利用一些距离度量方法来评估行人之间的相似性. 随着 Krizhevsky 赢得了 ILSVRC12[5] 的比赛, 基于深度学习的方法得以流行. 深度学习的方法主要集中于 3 个方面: 1) 通过设计卷积神经网络更好地学习到行人的特征; 2) 利用损失函数更好地度量行人相似度; 3) 通过数据增强让网络更加鲁棒, 使网络可以忽略一些和行人类别无关的特征.

Gray 等[6] 为了考虑到空间信息, 首先将图像按水平方向划分为多个矩形, 之后在每个矩形内, 利用颜色特征中的 RGB、HSV、YCbCr, 以及选择 21 个 Gabor、Schmid 滤波核来获得纹理特征. 最后将得到的每个水平条特征拼接在一起, 作为最后行

人的特征表示.

Yang 等[7] 提出了一种新的语义特征显著性 Color Name 特征, 该特征不同于传统的颜色直方图, 它通过将颜色量化, 保证每一个像素的颜色通道以较大的概率划分到量化的颜色区间, 即对应的 Color name 中.

2012 年 Köstinger 等[8] 提出经典的基于马氏距离度量的行人再识别算法 KISSME (Keep it simple and straightforward metric).

Zheng 等[9] 利用一个孪生网络[10], 结合分类问题与验证问题, 一次输入一对行人图片, 对于输入的一对行人图片, 网络一方面要预测两幅图片中行人各自的 ID, 另一方面要判断输入的两幅图片中的行人是否为属于同一行人. 在分类问题中, 他们使用 SoftMax 损失进行行人类别分类. 在验证问题中, 利用一个二维 SoftMax 损失进行一个二分类.

Zhang 等[11] 提出了一种端到端的方法 AlignedReID, 让网络自动地去学习人体对齐. 在 AlignedReID 中, 深度卷积神经网络不仅提取全局特征, 同时也对各局部提取局部信息, 在提取局部信息时采用动态匹配的方法选取最短路径, 从而进行行人对齐, 在训练时, 最短路径长度被加入到损失函数, 辅助学习行人的整体特征.

Zhao 等[12] 提出了一种基于人体关节点对人体进行区域划分的网络 (Spindle net), 首先定位人体的 14 个关节点, 通过区域提取网络来产生 7 个身体区域, 再通过 FEN (Feature extraction net) 特征提取网络和 FFN (Feature fusion net) 特征融合网络以身体区域为基础进行人体特征提取与融合.

Dai 等[13] 提出了一种批特征擦除 BFE (Batch feature erasing) 方法, 对于一个批量的特征图, 随机遮挡住同样的一块区域, 强迫网络在剩余的区域里面去学一些细节的特征. 这样训练得到的网络不会太过于关注那些显而易见的全局特征.

Zhong 等[14] 通过引入 Camera style adaptation 来解决相机差异导致的行人图片变化 (光线、角度等) 的问题. 作者首先利用 CycleGAN[15] 实现不同相机风格的转化, 在得到不同相机风格下的图片后, 将这些生成的图片放入网络中进行训练, 其中原始图像利用 SoftMax 损失进行有监督的训练, 生成图像利用 LSR (Label smoothing regularization) 损失进行训练. LSR 损失用于解决生成图像产生较多噪音的问题. 通过在训练数据中增加相机风格图片, 一方面增加了训练集数据量, 另一方面通过增加各个相机风格图片, 使得网络能够集中学习与相机无关的特征.

1.2 跨模态行人再识别方法

跨模态行人再识别的方法目前集中于深度学习的方法. 包括通过设计卷积神经网络来更好地学习跨模态行人的特征以及利用损失函数来更好地度量不同模态的行人之间的相似度.

2017 年, Wu 等[16] 提出了一种基于 Deep zero-padding 的跨模态行人再识别方法, 并且建立了一个大规模跨模态行人再识别数据集 SUSU-MM01. 作者对输入的 RGB 图和 IR 图在通道上进行了填充. RGB 图先转换为第 1 通道的灰度图, 之后在第 2 通道填充大小与灰度图一样的全 0 值. 对 IR 图, 在第 1 通道填充大小与 IR 图一样的全 0 值. 接着将填充后的 RGB 图和 IR 图统一的放入网络中进行训练, 通过 SoftMax 损失对行人标签进行有监督的训练.

Ye 等[17] 提出 BDTR (Bi-directional dual-constrained top-ranking) 方法来解决跨模态行人再识别. 作者通过一个孪生网络对 RGB 图片和 IR 图片分别进行特征提取, 利用 SoftMax 损失和提出的双向排序损失 (Bi-directional ranking loss) 进行有监督的训练. 双向排序损失包括跨模态约束 (Cross-modality top-ranking constraint) 和模态内约束 (Intra-modality top-ranking constraint).

Dai 等[18] 提出了 cmGAN (Cross-modality generative adversarial network) 方法, 该方法同样使用了类似于 BDTR 中的跨模态约束损失来保证跨模态图像的负样本对距离大于跨模态图像的正样本对距离, 另外, 利用 SoftMax 损失对行人 ID 进行有监督的训练. 除此之外, 结合生成对抗网络的对抗训练的思想, 在判别器部分, 用一个二分类来区分图像是 RGB 图还是 IR 图.

Lin 等[19] 提出了 HPILN (Hard pentaplet and identity loss network) 方法, 该方法对现有的单个模态的行人再识别模型进行了改进, 使其更适用于跨模态场景, 并提出一个新型损失函数: Hard 五元组损失 (Hard pentapelt loss), 使得网络可以同时处理模态内和模态间变化, 再结合身份损失函数 (Identity loss) 来提高改进后的模型的性能.

2 基于自注意力模态融合网络的跨模态行人再识别方法

跨模态行人再识别和传统的行人再识别相比, 增加了相同行人不同模态的变化. 为了减轻跨模态行人再识别中由于跨模态数据导致的问题, 本文首先利用 CycleGAN[15] 对于每一幅图片生成其对应

跨模态下的图片. 如果原始图片是 RGB 图, 则 CycleGAN 生成 IR 图; 如果原始图片是 IR 图, 则 CycleGAN 生成 RGB 图. 之后利用跨模态学习网络将原始数据和生成的跨模态数据加入到基本的分类网络中进行训练, 这样跨模态学习网络即可同时利用原始数据以及经过 CycleGAN 生成的跨模态数据. 对于每一幅图片, 为了将原始图片与 CycleGAN 生成的跨模态数据进行区分以及特征选择, 本文针对每一种数据, 分别设计了一个自注意力模块进行行人特征的筛选. 接着将经过自注意力模块后的原始特征和跨模态图片特征经过 Max 层进行融合, 最后原始图片特征以及融合后的特征利用 SoftMax 损失进行有监督的训练, CycleGAN 生成的跨模态图片特征利用 LSR 损失[20] 进行训练. 自注意力模态融合网络的结构图如图 3 所示.

2.1 跨模态图像生成

生成对抗网络 (Generative adversarial network, GAN)[21−22] 自 2014 年由 Goodfellow 等提出后, 越来越受到学术界和工业界的重视. 其中, GAN 在图像生成上取得了巨大的成功, 这取决于 GAN 在博弈下不断提高建模能力, 最终实现以假乱真的图像生成. 图像到图像的转换可分为有监督 (如 cGAN[23], pix2pix[24]) 和无监督 (如 CycleGAN[15], DualGAN[25]) 两大类.

针对本文的跨模态应用场景, 我们没有成对的样本数据作为输入图像, 所以无监督的生成对抗网络更适用; 其次, 尽管 CycleGAN 和 DualGAN 具有相同的模型结构, 但它们对生成器使用不同的实现方法. CycleGAN 使用卷积架构的生成器结构, 而 DualGAN 遵循 U-Net 结构; CycleGAN 重在解决非配对图像转换问题, 而 DualGAN 重在解决如何避免模型崩溃问题. 经过以上综合分析, CycleGAN 适合完成风格迁移任务且是无监督的, 因此更适用于我们的网络.

为了学习到跨模态的信息, 本文首先利用 CycleGAN 生成跨模态的数据. CycleGAN 可以将两个域的图像进行相互转换, 并且 CycleGAN 的输入是任意的两幅图片, 不需要它们成对出现. 因此, 可以直接利用 CycleGAN 实现跨模态行人再识别中的数据模态转换. CycleGAN 的网络结构如图 4 所示.

假设有来自两个属于不同数据域的集合, 记为 A, B. CycleGAN 由两个判别器 D (分别记为 D_A、D_B) 和两个生成器 G (分别记为 G_{AB}、G_{BA}) 组成. 其中 G_{AB} 用来将 A 域的图像转换到 B 域, G_{BA} 用来将 B 域的图像转换到 A 域. D_A 判断输入图片是否是真实的图片, 即图片是 A 域的原始图片还是 G_{BA} 转换后的生成图片. 其目标是将生成模型 G_{BA} 产生的 “假”图片和训练集 A 域中 “真”图片进行区分. 同样, D_B 用来判断图片是 B 域的原始图片还

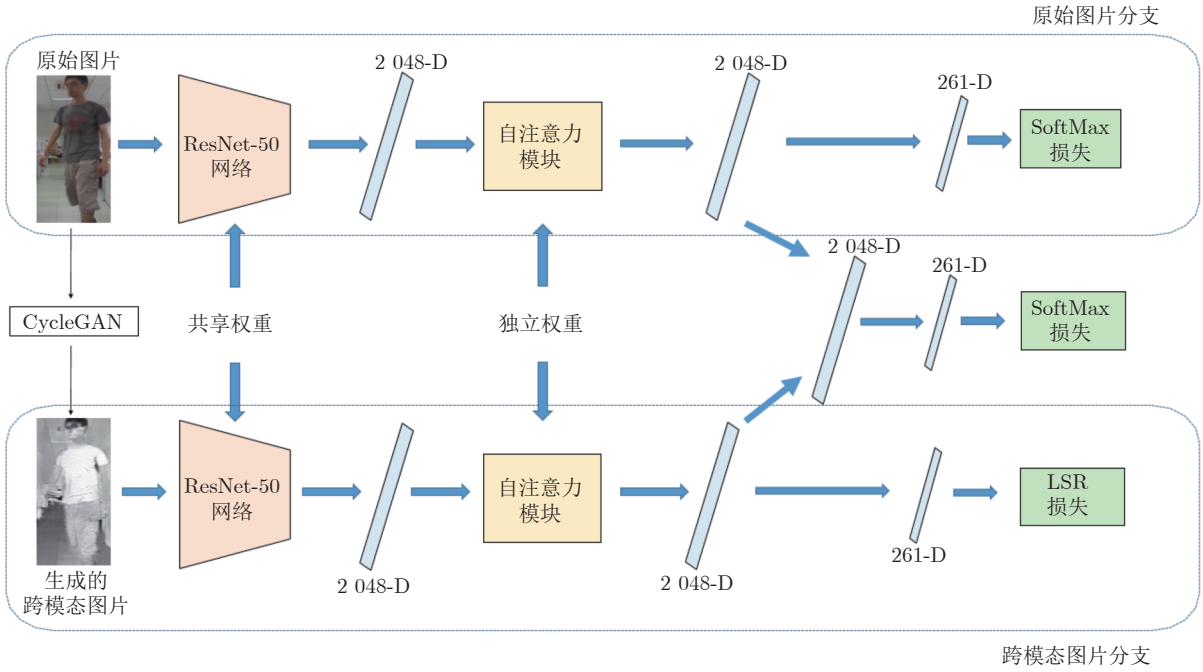

图 3 自注意力模态融合网络

Fig. 3 Self-attention cross-modality fusion network

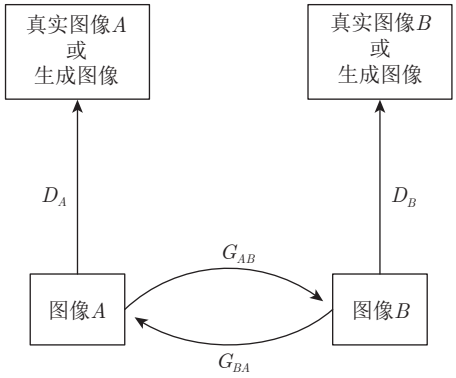

图 4　CycleGAN 网络示意图

Fig. 4　Structure of CycleGAN network

是 G_{AB} 转换后的生成图片. 其目标是将生成模型 G_{AB} 产生的 “假”图片和训练集 B 域中 “真”图片进行区分.

2.2　跨模态学习网络

本文将所有的 RGB 相机采集到的图像作为域 A, 所有的红外相机采集到的 IR 图像作为域 B. 图像统一缩放为 256×128 像素. 将训练数据中的 A 域和 B 域送入 CycleGAN 中进行训练. 当训练完成后, 利用 G_{AB} 将原始的 RGB 图像转换为 IR 域风格图像, 利用 G_{BA} 将原始的 IR 图像转换为 RGB 域风格图像, 这样对于同一幅图像, 既有原始域的图像, 也有风格转换即跨模态的图像. 经过 CycleGAN 生成的跨模态图如图 5 所示. 其中第 1 行为数据集中的 RGB 图, 第 2 行为利用 CycleGAN 生成的对应的跨模态 IR 图, 第 3 行为数据集中的 IR 图, 第 4 行为利用 CycleGAN 生成的对应的跨模态 RGB 图. 同一列为相同的行人, 不同列对应不同行人. 可以看出, 利用 CycleGAN 可以大致地实现数据的跨模态变化.

跨模态学习网络的设计参照 Zhong 等[14] 设计的网络. 该网络由一对共享权重的 ResNet-50 组成. 在得到了两种模态图像后, 本节将原始的数据和生成的跨模态数据都加入到 ResNet-50[26] 网络中进行训练. 跨模态学习网络的输入和一般的分类网络不同, 它的输入为一对图像, 包括原始图像和 CycleGAN 生成的跨模态图, 跨模态学习网络每次输入的生成图像是由原始图像生成的跨模态图. 由于生成图像是由原始图像变换过来, 所以该生成图像的标签理想情况下应该和原始图像标签一致, 因此在训练跨模态生成图时可以和原始图像一样, 可以利用 SoftMax 损失进行有监督的训练. SoftMax 损失的计算如式 (1) 所示.

图 5　利用 CycleGAN 生成的跨模态图像

Fig. 5　Generated cross-modality images using CycleGAN

$$L_{\text{Cross}} = -\sum_{k=1}^{K} q(k) \lg p(k) \tag{1}$$

式中, L_{Cross} 表示 SoftMax 损失; K 为类别数; $q(k)$ 表示真实标签的 One hot 形式, 即真实数据分布; $p(k)$ 表示预测的结果.

但是, 在观察生成的跨模态图时, 发现生成的跨模态图大多具有很大的噪声, 尤其是当 IR 图像到 RGB 图像的转换时. 如图 6 所示, 其中第 1 行为原始的 RGB 图; 第 2 行为利用 CycleGAN 生成的对应的跨模态 IR 图; 第 3 行为原始的 IR 图; 第 4 行为利用 CycleGAN 生成的对应的跨模态 RGB 图. 同一列为相同的行人, 不同列对应不同行人. 从中可以看出, 生成的图像一般很难和原始图像用一个标签来区分.

本文针对跨模态行人再识别中数据集的模态变化问题, 提出了一种自注意力模态融合网络. 采用 CycleGAN 进行跨模态图像的生成, 并在 ResNet50

图 6 包含较多噪声的跨模态转换后的图像

Fig. 6 Generated cross-modality images with more noise

网络的基础上加入了自注意力模块和模态融合模块. 通过对网络中的不同模块进行组合对比实验, 证明了本节提出的每一个模块的有效性. 另外通过在 SYSU-MM01 数据集上的实验, 也证明了本文提出的方法与其他跨模态方法相比有一定程度的提升. 与其他跨模态行人再识别方法相比, 本文不仅在网络结构上进行了改进, 同时在数据层面进行了创新. 我们首次将 CycleGAN 用于跨模态行人再识别图像生成从而实现数据的跨模态变化. 但目前本文方法跨模态生成的图像质量较差, 有一定的噪声. 为了克服以上缺陷, 在今后的工作中将重点解决此问题, 从而更好地解决跨模态行人再识别问题.

针对上述问题, 对于 CycleGAN 生成的跨模态图, 本文利用 LSR 损失来进行训练. 一般的分类损失函数, 如 SoftMax 损失, 对图像的标签会编辑成 One hot 形式, 如式 (2) 所示. LSR 损失考虑到数据的过拟合, 在给定图像标签时, LSR 给定 Ground-truth 类一个比较大的值, 剩余的类标签给定一个比较小的值, 如式 (3) 所示, 将 LSR 的数据标签代入 SoftMax 损失 (式 (1)) 中, 即得到 LSR 的计算式, 如式 (4) 所示.

$$q(k)=\begin{cases}1, & k=y\\ 0, & k\neq y\end{cases} \tag{2}$$

式中, $q(k)$ 表示 SoftMax loss 中行人类别的 One hot 编码; y 表示真实数据标签.

$$q_{\mathrm{LSR}}(k)=\begin{cases}1-\varepsilon+\dfrac{\varepsilon}{K}, & k=y\\ \dfrac{\varepsilon}{K}, & k\neq y\end{cases} \tag{3}$$

式中, $q_{\mathrm{LSR}}(k)$ 表示 LSR 损失中行人类别的编码; ε 表示平滑参数; K 表示行人类别数; y 表示真实数据标签.

$$L_{\mathrm{LSR}}=-(1-\varepsilon)\lg p(k)-\frac{\varepsilon}{K}\sum_{k=1}^{K}\lg p(k) \tag{4}$$

式中, L_{LSR} 表示 LSR 损失, ε 表示平滑参数, 本节中设定 ε 为 0.1, $p(k)$ 表示预测的结果, K 表示行人类别数.

跨模态学习网络同时学习了原始图像以及相对应的跨模态图像的特征. 同时, 由于对同一幅图网络得到了两种模态信息, 数据量上有一定的提升, 可以看成是进行了数据增广. 除此之外, 网络对同一幅图同时考虑到了两种模态信息, 因此, 跨模态学习网络同时增强了对于模态无关特征的学习.

2.3 自注意力模块

在第 2.2 节中的跨模态学习网络, 虽然同时输入了两种模态图像, 但是除了在最后损失函数的时候进行区分外, 网络本身对于原始图像和跨模态图像的处理完全一致. 这样通过数据增广的方式在一定程度上虽然可以学习到一些模态无关的特征, 但是不同模态之间缺少交互, 在训练过程中两种模态之间单独地进行监督训练. 卷积神经网络通过在局部感受野上进行卷积操作来融合空间和通道信息, 而自注意力模块本质上引入了对输入的动态适应性, 这有助于增强特征区分能力, 提高行人再识别的性能. 因此, 针对上述问题, 本文在跨模态学习网络的基础上构建了一个自注意力模块, 该模块通过自注意力机制将原始图像和 CycleGAN 生成的图像进行区分, 自动地对第 2.2 节中产生的不同模态的特征在通道层面进行筛选. 该模块增加在跨模态学习网络的 2 048 维特征层和最后一层 261 维 (与训练数据集中行人数一致) 全连接层之间. 它的输入是经过跨模态学习网络产生的两个 2 048 维特

征, 经过自注意力模块后, 输出依然为两个 2 048 维特征, 该特征维度和跨模态学习网络的输出维度一致, 但是对不同模态的特征进行了筛选.

自注意力模块的设计参照 SENet[27] 中 SE (Squeeze-and-excitation) 模块. 由于自注意力模块是直接在 ResNet-50 全局平局池化后的特征通道上进行特征选择, 因此和 SE 模块不同, 自注意力模块不需要额外使用全局平均池化做一个 Squeeze 操作. 剩余 Excitation 操作和 SE 模块保持一致. 自注意力模块包括两个全连接层、一个 ReLU[28] 激活函数和一个 Sigmoid[29] 激活函数. 自注意力模块使用两个全连接层去构造特征通道间的相关性. 首先, 第 1 个全连接层将特征维度降低到输入的 k 分之一. 在本节中设定 k 和 SENet 中的一致, 为 1/16. 降维后再经过 ReLU 激活函数激活, 之后再通过一个全连接层恢复到原来的输入特征维度. 通过这样的设计增加了自注意力模块的非线性, 可以更好地拟合复杂的特征空间. 另外通过这样构造的两层全连接层极大地减少了参数量和计算量. 之后通过一个 Sigmoid 激活函数获得最后的特征权重, 由于经过 Sigmoid 激活, 得到的权重值在 0 至 1 之间. 最后将得到的权重和原始的特征按元素相乘, 这样就实现了自注意力模块. 自注意力模块的网络结构图如图 7 所示.

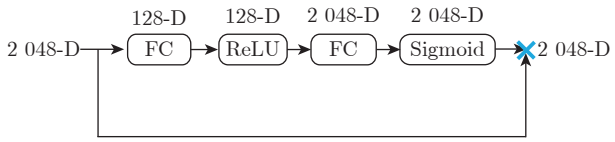

图 7　自注意力模块示意图

Fig. 7　Structure of self-attention model

2.4　模态融合模块

在利用第 2.2 节中的跨模态学习网络进行行人再识别测评时, 仅仅输入原始图像, 测试集的生成图像并没有得到充分利用. 针对该问题, 本节提出利用模态融合模块将两种筛选后的特征进行融合, 融合后的结果再送入到全连接层, 最后用 SoftMax 损失进行有监督的训练.

模态融合模块的目的是将原始图像和 CycleGAN 生成的图像进行融合. 由于 CycleGAN 生成的图像相对于原始图像是跨模态的, 即原始图像如果是 RGB 图, CycleGAN 生成的图是 IR 图, 反之如果原始图像是 IR 图, CycleGAN 生成的图是 RGB 图. 这两种图像应该具有互补性. 在一定的条件下, 通过 RGB 图可以获得丰富的颜色特征, 通过 IR 图可以获得丰富的纹理特征. 因此, 在本节利用模态融合网络可以将原始图像以及跨模态图像中对于分类比较有用的特征进行保留. 模态融合模块通过一个 Max 层完成. 将经过自注意力模块的原始图像特征和 CycleGAN 生成图像的特征经过 Max 层进行融合. 融合后的特征再连接到共享的全连接层上, 最后进行有监督的训练.

3　实验结果与分析

为了评价自注意力模态融合网络在跨模态行人再识别中的效果, 本节在一个常用的跨模态行人再识别数据集 SYSU-MM01[16] 上进行实验. 评价指标选择了行人再识别中常用的 CMC 曲线 (Cumulative matching curve) 和 mAP (mean average precision).

3.1　数据集与评价指标

SYSU-MM01 是中山大学采集的一个跨模态行人再识别数据集. 它包括 4 个 RGB 相机和两个 IR 相机. 其中 cam1 与 cam2 为拍摄到的 Indoor 场景下的 RGB 图像, cam3 为 Indoor 场景下的 IR 图像, 且与 cam2 是同一个场景; cam4 与 cam5 为 Outdoor 场景下的 RGB 图像, cam6 为 Outdoor 场景下的 IR 图像. SYSU-MM01 总共有 491 个不同行人, 总共包括 287 628 幅 RGB 图像, 15 792 幅 IR 图像.

在测试的时候, 该数据集中测试集的所有 IR 图像作为 Probe, 所有的 RGB 图像作为 Gallery. 有两种评价模式, 一种是 All-search 模式, 另一种是 Indoor-search 模式. 除此之外, 在每种模式下, 分别采用 Single-shot 测评和 Multi-shot 测评. 在 Single-shot 测评时, 在测试集中的每一个行人, Gallery 集合中随机选取一个与该行人类别相同的 RGB 图片构成 Gallery 集, 所有的 Probe 图像构成 Probe 集. 在 Multi-shot 测评时, 对于测试集中的每一个行人, Gallery 集合中随机选取 10 个与该行人类别相同的 RGB 图片构成 Gallery 集, 所有的 Probe 图像构成 Probe 集.

在该数据集上测评时, 使用 CMC 曲线和 mAP 来进行测评. 在测评时, 利用上述的方法构造 Probe 和 Gallery. 计算 CMC 曲线和 mAP 的方法和传统的行人再识别方法一致. 但是, 考虑到该数据集下 cam2 和 cam3 是在同一个地方采集, 而行人再识别的研究重点是跨摄像头, 因此, 在评价算法时, 在匹配 cam2 的 Probe 时, 会忽略 cam3 中的 Gallery. 对于上述的每一种测评, 包括 All-search 下的

Single-shot 测评和 Multi-shot 测评以及 Indoor-search 下的 Single-shot 测评和 Multi-shot 测评, 本文都重复了 10 次实验并计算 10 次的平局值.

3.2 实现细节

我们使用 Pytorch[30] 来实现本文中的自注意力模态融合网络. 在训练过程中, 跨模态学习网络首先加载了在 ImageNet 上预训练的 ResNet-50 网络的参数. 我们使用 AMSGrad[31] 来训练网络. 给定权重衰减 (Weight decay) 为 5×10^{-4} 来减轻网络过拟合.

训练过程分为两个阶段. 第 1 阶段是训练第 2.2 节中的跨模态学习网络. 在这一阶段中, 训练 Batch size 设定为 32, 总共训练 60 轮, 初始学习率为 3×10^{-4}, 学习率每过 20 轮变为原始的 1/10. 第 2 阶段训练整个自注意力模态融合网络, 加载第 1 阶段训练好的跨模态学习网络参数, 之后训练自注意力模态融合网络, 训练 Batch size 设定为 32, 总共训练 60 轮, 初始学习率为 3×10^{-4}, 学习率每过 20 轮变为原始的 1/10.

3.3 实验结果与分析

3.3.1 不同模块组合对比实验

为了测试自注意力模块融合网络中每一个模块的有效性. 本节总共构建了 5 个网络. 第 1 个是一般的分类网络, 用作跨模态行人再识别的 Baseline 网络, 该网络由一个 ResNet-50 组成, 这里将其命名为 “Baseline”; 第 2 个是第 2.2 节中构建的跨模态学习网络; 第 3 个是在跨模态学习网络的基础上加入自注意力模块, 命名为 “跨模态 + 自注意力”; 第 4 个是在跨模态学习网络的基础上加入融合模块, 命名为 “跨模态 + 模态融合”; 第 5 个是在跨模态学习网络的基础上加入融合模块以及自注意力模块, 即本文中的自注意力融合网络. 这 5 组网络在 SYSU-MM01 的实验结果如表 1 和表 2 所示, 表中汇集了 CMC 曲线中的 Rank 1、Rank 10、Rank 20 以及 mAP 的实验结果.

从表 1 和表 2 可以看出, 与 Baseline 相比, 在引入了 CycleGAN 生成的图像并利用跨模态学习网络同时训练原始图像和跨模态图像时, 在 SYSU-MM01 数据集上的成绩有显著的提升. 在 All-search 模式下, Single-shot 和 Mulit-shot 的 Rank 1 分别提升了 3.47% 和 4.77%. 在 Indoor-search 模式下, Single-shot 和 Mulit-shot 的 Rank 1 分别提升了 5.04% 和 5.03%. 这组对比实验说明了在第 2.2 节中提出的跨模态学习网络的有效性. 跨模态学习网络和 Baseline 相比, 同时利用了原始图像和生成的跨模态图像.

表 1 各模块在 SYSU-MM01 All-search 模式下的实验结果

Table 1 Experimental results of each module in SYSU-MM01 dataset and All-search mode

方法	All-search							
	Single-shot				Multi-shot			
	Rank 1	Rank 10	Rank 20	mAP	Rank 1	Rank 10	Rank 20	mAP
Baseline	27.36	71.95	84.58	28.53	32.48	78.34	88.93	23.17
跨模态学习	30.83	72.35	84.07	31.45	37.25	80.58	90.22	25.48
跨模态 + 自注意力	31.3	73.34	84.78	31.72	37.98	81.76	91.05	25.39
跨模态 + 模态融合	31.85	74.38	85.66	32.49	38.65	81.74	**91.25**	26.46
自注意力模态融合	**33.31**	**74.51**	**85.79**	**33.18**	**39.71**	**82**	91.14	**26.89**

表 2 各模块在 SYSU-MM01 Indoor-search 模式下的实验结果

Table 2 Experimental results of each module in SYSU-MM01 dataset and Indoor-search mode

方法	Indoor-search							
	Single-shot				Multi-shot			
	Rank 1	Rank 10	Rank 20	mAP	Rank 1	Rank 10	Rank 20	mAP
Baseline	32.17	81.3	**92.26**	42.76	38.95	85.29	93.62	33.73
跨模态学习	37.21	80.81	90.29	47.06	43.98	86.01	93.37	37.09
跨模态 + 自注意力	36.55	80.32	90.41	46.42	44.89	85.31	94.18	36.43
跨模态 + 模态融合	37.63	**81.75**	91.48	47.73	44.82	**87.26**	**94.97**	**38.07**
自注意力模态融合	**38.09**	81.68	90.61	**47.86**	**45.8**	86.72	93.86	37.95

对比自注意力模态融合网络和第 2.2 节中的跨模态学习网络, 发现自注意力模态融合网络成绩有更近一步的提升. 在 All-search 模式下, Single-shot 和 Mulit-shot 的 Rank 1 分别提升了 2.48% 和 2.46%. 在 Indoor-search 模式下, Single-shot 和 Mulit-shot 的 Rank 1 分别提升了 0.88% 和 1.82%. 这组对比实验说明了本文提出的自注意力模态融合网络的有效性. 最后, 单独比较自注意力模态融合网络和 “跨模态+自注意力”以及 “跨模态+模态融合”, 发现由于生成图像存在很大的噪声, 对自注意力模块造成了一定程度的影响. 从而导致在 Indoor-search 和 Multi-shot 模式下, “自注意力模态融合”的 mAP 比起 “跨模态 + 模态融合”下降了 0.12%, 如何对生成的图像降噪是今后要解决的问题之一. 不过, 从总体来看, 两个模块共同使用比单独使用它们中的任一个模块都要有效.

我们参照 SENet[27] 中对网络时间复杂度的分析方法, 计算了在测试时加入各个模块后网络的 GFLOPs (Giga floating-point operatiuns per second) 和参数量, 如表 3 所示. 其中, 前三个方法的输入是一幅大小为 256×128 像素的图像, “跨模态 + 模态融合”网络和 “自注意力模态融合”网络的输入是一幅大小为 256×128 像素的图像和一幅生成的相同大小的跨模态图像. 由表 3 可知, 跨模态学习网络与 Baseline 相比, GFLOPs 和参数量都相同; 加入自注意力模块后, GFLOPs 增加了 0.001048576, 参数量增加了 4.12%; 由于输入是两幅图, “跨模态 + 模态融合”网络 GFLOPs 是 Baseline 的两倍, 由于 Max 操作没有新增参数, 所以参数量没有发生变化. “自注意力模态融合”网络与 Baseline 相比, GFLOPs 增加了 2.706867200, 参数量增加了 6.18%. 可见自注意力模块对 GFLOPs 的影响微乎其微, GFLOPs 的增加主要来源于输入的增加.

3.3.2　和跨模态行人再识别 State-of-the-arts 对比实验

我们在 SYSU-MM01 数据集上和跨模态行人再识别 State-of-the-arts 进行了对比. 其中 “HOG + Euclidean”是在 RGB-RGB 匹配的行人再识别问题中利用模式识别方法解决, 手工特征选择 HOG[32] 特征, 距离度量利用欧氏距离度量; “LOMO + KISSME”同样也是利用传统的模式识别方法, 手工特征选择 LOMO[33] 特征, 距离度量算法利用 KISSME[8]; “Zero-padding”[16] 方法属于深度学习方法中的基于深度特征学习法, 该方法将三通道的 RGB 图转换为一通道的灰度图, 之后在第 2 通道进行零值填充, 将 IR 图直接在第 1 通道进行零值填充, 之后将填充后的 RGB 图和 IR 图统一放入网络中, 利用 SoftMax 损失进行训练; BDTR[17] 属于深度学习方法中的基于距离度量学习法, 该方法通过一个孪生网络对 RGB 图片和 IR 图片分别进行特征提取, 利用 SoftMax 损失和双向排序损失进行有监督的训练; cmGAN[18] 属于深度学习方法中的基于距离度量学习法, 该方法使用三元组损失来约束跨模态样本距离, 保证跨模态图像的负样本对距离大于跨模态图像的正样本对距离, 同时利用 SoftMax 损失对行人 ID 进行有监督的训练. 另外结合 GAN 网络对抗训练的思想, 在判别器部分用一个二分类来区分图像是 RGB 图还是 IR 图. 与上述 4 个方法对比的实验结果如表 4 和表 5 所示.

从表 4 和表 5 可以看出, 基于深度学习的跨模态行人再识别方法要远远好于传统的模式识别方法. 另外, 由于跨模态行人再识别目前的研究工作较少, 早期的 Zero-padding 利用的基网络为 ResNet-6, BDTR 利用的基网络为 AlexNet[5]. 本文中利用的基网络和 cmGAN 方法中的基网络一致, 为 ResNet-50. ResNet-50 也是 RGB-RGB 行人再识别中最常用的基网络. 从实验结果看, 本文中提出的自注意力模态融合网络相较于上述方法成绩有一个比较大的提升. 在 All-search 模式下, Single-shot 的 Rank 1 相比于 Zero-padding、BDTR 和 cmGAN 分别提升 18.51%、16.3% 和 6.04%. Multi-shot 的 Rank 1 相比于 Zero-padding 和 cmGAN

表 3　加入各模块后的 GFLOPs 和参数量
Table 3　GFLOPs and parameters after joining each module

方法	GFLOPs	GFLOPs相比于Baseline的变化	参数量	参数量相比于Baseline的变化
Baseline	2.702772224	–	25 557 032	–
跨模态学习	2.702772224	0	25 557 032	0
跨模态 + 自注意力	2.7038208	0.001048576	26 609 960	+1 052 928 (4.12%)
跨模态 + 模态融合	5.405544448	2.702772224	25 557 032	0
自注意力模态融合	5.409639424	2.7068672	27 136 424	+1 579 392 (6.18%)

表 4 在 SYSU-MM01 All-search 模式下和跨模态行人再识别的对比实验

Table 4 Comparative experiments between our method and others in SYSU-MM01 dataset and All-search mode

方法	All-search							
	Single-shot				Multi-shot			
	Rank 1	Rank 10	Rank 20	mAP	Rank 1	Rank 10	Rank 20	mAP
HOG + Euclidean	2.76	18.25	31.91	4.24	3.82	22.77	37.63	2.16
Zero-padding	14.8	54.12	71.33	15.95	19.13	61.4	78.41	10.89
BDTR	17.01	55.43	71.96	19.66	–	–	–	–
cmGAN	26.97	67.51	80.56	27.8	31.49	72.74	85.01	22.27
Baseline (本文方法)	27.36	71.95	84.58	28.53	32.48	78.34	88.93	23.17
跨模态学习网络 (本文方法)	30.83	72.35	84.07	31.45	37.25	80.58	90.22	25.48
自注意力模态融合 (本文方法)	**33.31**	**74.51**	**85.79**	**33.18**	**39.71**	**82**	**91.14**	**26.89**

表 5 在 SYSU-MM01 Indoor-search 模式下和跨模态行人再识别的对比实验

Table 5 Comparative experiments between our method and others in SYSU-MM01 dataset and Indoor-search mode

方法	Indoor-search							
	Single-shot				Multi-shot			
	Rank 1	Rank 10	Rank 20	mAP	Rank 1	Rank 10	Rank 20	mAP
HOG + Euclidean	3.22	24.68	44.52	7.25	4.75	29.06	49.38	3.51
Zero-padding	20.58	68.38	85.79	26.92	24.43	75.86	91.32	18.64
cmGAN	31.63	77.23	89.18	42.19	37	80.94	92.11	32.76
Baseline (本文方法)	32.17	81.3	**92.26**	42.76	38.95	85.29	93.62	33.73
跨模态学习网络 (本文方法)	37.21	80.81	90.29	47.06	43.98	86.01	93.37	37.09
自注意力模态融合 (本文方法)	**38.09**	**81.68**	90.61	**47.86**	**45.8**	**86.72**	**93.86**	**37.95**

分别提升 20.4% 和 8.22%. 在 Indoor-search 模式下, Single-shot 的 Rank 1 相比于 Zero-padding 和 cmGAN 分别提升 17.51% 和 6.46%. Multi-shot 的 Rank 1 相比于 Zero-padding 和 cmGAN 分别提升 21.37% 和 8.8%. 可以看出, 本文提出的自注意力模态融合网络在 SYSU-MM01 数据集上已经超过了现有的跨模态行人再识别方法.

4 结束语

跨模态行人再识别与传统的行人再识别相比, 增加了相同行人不同模态的变化. 为了解决跨模态问题, 本文提出了一种自注意力模态融合网络. 首先利用 CycleGAN 生成原始图像的跨模态图像, 之后利用跨模态学习网络将两个模态的图片都加入网络进行训练. 接着利用自注意力模块对原始图像和 CycleGAN 生成的图像分别进行特征筛选, 最后利用模态融合模块将两种模态的图片特征融合作为最后的行人再识别中行人的特征表示. 在 SYSU-MM01 数据集上的实验结果证明了本文提出的方法和其他跨模态方法相比有一定程度的提升. 本文首次将 CycleGAN 用于跨模态行人再识别图像生成, 实现数据的跨模态变化. 不仅在网络结构上进行了改进, 同时在数据层面进行了创新. 在今后的工作中将致力于提升跨模态生成的图像质量从而更好地解决跨模态行人再识别问题.

References

1 Li You-Jiao, Zhuo Li, Zhang Jing, Li Jia-Feng, Zhang Hui. A survey of person re-identiflcation. *Acta Automatica Sinica*, 2018, **44**(9): 1554−1568
(李幼蛟, 卓力, 张菁, 李嘉锋, 张辉. 行人再识别技术综述. 自动化学报, 2018, **44**(9): 1554−1568)

2 Wu Yan-Cheng, Chen Hong-Chang, Li Shao-Mei, Gao Chao. Person re-identiflcation using attribute priori distribution. *Acta Automatica Sinica*, 2019, **45**(5): 953−964
(吴彦丞, 陈鸿昶, 李邵梅, 高超. 基于行人属性先验分布的行人再识别. 自动化学报, 2019, **45**(5): 953−964)

3 Zhang L, Ma B P, Li G R, Huang Q M, Tian Q. Generalized semisupervised and structured subspace learning for cross-modal retrieval. *IEEE Transactions on Multimedia*, 2018, **20**: 128−141

4 Zhang L, Ma B P, Li G R, Huang Q M, Tian Q. PL-ranking: A novel ranking method for cross-modal retrieval. In: Proceedings of the 24th ACM on Multimedia Conference. Amsterdam, the Netherlamds: ACM, 2016. 1355−1364

5 Krizhevsky A, Sutskever I, Hinton G. ImageNet classifcation with deep convolutional neural networks. *Communications of the ACM*, 2012, **60**: 84−90

6 Gray D, Tao H. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: Proceedings of the 10th European Conference on Computer Vision, Marseille, France: Springer, 2008. 262−275

7 Yang Y, Yang J M, Yan J J, Liao S C, Yi D, Li S Z. Salient color names for person re-identifcation. In: Proceedings of the 2014 Eruopeam Computer Vision. Zuich, Switzerland: Springer, 2014. Part I : 536−551

8 Köstinger M, Hirzer M, Wohlhart P, Roth P M, Bischof H. Large scale metric learning from equivalence constraints. In: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence, USA: IEEE, 2012. 2288−2295

9 Zheng Z D, Zheng L, Yang Y. A discriminatively learned CNN embedding for person reidentifcation. *ACM Transactions on Multimedia Computing Communications and Applications*, 2016, **14**(1): 13-1−13-20

10 Bromley J, Bentz J W, Bottoul, Guyon I, Lecun Y, Moore C, et al. Signature verifcation using a "siamese" time delay neural network. *International Journal of Pattern Recognition and Artifcial Intelligence*, 1993, **7**(4): 669−688

11 Zhang X, Luo H, Fan X, Xiang W L, Sun Y X, Xiao Q Q, Jiang W, Zhang C, Sun J. AlignedReID: Surpassing humanlevel performance in person re-identifcation. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Hawaii, USA: IEEE, 2017.

12 Zhao H Y, Tian M Q, Sun S Y, Shao J, Yan J J, Yi S, Wang X G, Tang X O. Spindle net: Person re-identifcation with human body region guided feature decomposition and fusion. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE, 2017, 907−915

13 Dai Z Z, Chen M Q, Zhu S Y, Tan P. Batch feature erasing for person re-identifcation and beyond. *Computer Research Repository*, 2018.

14 Zhong Z, Zheng L, Zheng Z D, Li S Z, Yang Y. Camera style adaptation for person re-identifcation. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake, USA: IEEE, 2018. 5157−5166

15 Zhu J Y, Park T, Isola P, ECfros A A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017. 2242−2251

16 Wu A C, Zheng W S, Yu H X, Gong S G, Lai J H. RGB-infrared crossmodality person re-identifcation. In: Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017. 5390−5399

17 Ye M, Wang Z, Lan X Y, Yuen P C. Visible thermal person reidentifcation via dual-constrained top-ranking. In: Proceedings of the 27th International Joint Conference on Artifcial Intelligence. Stockholm, Sweden: AAAI, 2018.1092−1099

18 Dai P Y, Ji R R, Wang H B, Wu Q, Huang Y Y. Cross-modality person re-identifcation with generative adversarial training. In: Proceedings of the 27th International Joint Conference on Artifcial Intelligence. Stockholm, Sweden: AAAI, 2018. 677−683

19 Lin J W, Li H. HPILN: A feature learning framework for crossmodality person re-identifcation. [Online], available: https://arxiv.org/abs/1906.03142, August 14, 2019

20 Szegedy C, Vanhoucke V, Iofie S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016. 2818−2826

21 Goodfellow I J, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. In: Advances in Neural Information Processing Systems, Berlin, Germany: Springer, 2014. 2672−2680

22 Lin Yi-Lun, Dai Xing-Yuan, Li Li, Wang Xiao, Wang Fei-Yue. The new frontier of AI research: Generative adversarial networks. *Acta Automatica Sinica*, 2018, **44**(5): 775−792
(林懿伦, 戴星原, 李力, 王晓, 王飞跃. 人工智能研究的新前线: 生成式对抗网络. 自动化学报, 2018, **44**(5): 775−792)

23 Mirza M P, Osindero S. Conditional generative adversarial nets. [Online], available: https://arxiv.org/abs/1411.1784, November 6, 2014

24 Isola P, Zhu J Y, Zhou T H, Efros A A. Image-to-image translation with conditional adversarial networks. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE, 2017. 5967−5976

25 Yi Z L, Zhang H, Tan P, Gong M L. DualGAN: unsupervised dual learning for image-to-image translation. In: Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017. 2868−2876

26 He K M, Zhang X Y, Ren S Q, Sun J. Deep residual learning for image recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016. 770−778

27 Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018. 7132−7141

28 Nair V, Hinton G E. Rectifled linear units improve restricted Boltzmann machines vinod nair. In: Proceedings of the 27th International Conference on International Conference on Machine Learning. Haifa, Israel: Omnipress, 2010. 807−814

29 Yin X Y, Goudriaan J W, Lantinga E A, Vos J C, Spiertz H L. A flexible sigmoid function of determinate growth. *Annals of Botany*, 2003, **91**(3): 361−371

30 Paszke A, Gross S, Chintala S, Chanan G, Yang E, DeVito Z, Lin Z M, Desmaison A, Antiga L, Lerer A. Automatic differentiation in PyTorch. In: Proceedings of the 31st Conference and Workshop on Neural Information Processing Systems. California, USA: NIPS, 2017.

31 Reddi S J, Kale S, Kumar S. On the convergence of Adam and Beyond. In: Proceedings of the 6th International Conference on Learning Representations. Vancouver, BC, Canada: ICLR, 2018.

32 Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego, CA, USA: IEEE, 2005. 886–893

33 Liao S C, Hu Y, Zhu X Y, Li S Z. Person re-identiflcation by local maximal occurrence representation and metric learning. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE, 2015. 2197–2206

杜 鹏 西安交通大学软件学院硕士研究生. 主要研究方向为行人再识别. E-mail: xjydupeng@163.com
(**DU Peng** Master student at the School of Software Engineering, Xi′an Jiaotong University. His main research interest is person re-identification.)

宋永红 西安交通大学人工智能学院研究员. 主要研究方向为图像与视频内容理解, 智能软件开发. 本文通信作者. E-mail: songyh@xjtu.edu.cn
(**SONG Yong-Hong** Researcher at the College of Artificial Intelligence, Xi′an Jiaotong University. Her research interest covers image and video content understanding, and intelligent software development. Corresponding author of this paper.)

张鑫瑶 西安交通大学软件学院硕士研究生. 主要研究方向为行人再识别. E-mail: xyzhangxy@stu.xjtu.edu.cn
(**ZHANG Xin-Yao** Master student at the School of Software Engineering, Xi′an Jiaotong University. Her main research interest is person re-identification.)