语义引导的遮挡行人再识别注意力网络

任雪娜[1,2]，张冬明[1,3]，包秀国[1,3]，李冰[4]

（1. 中国科学院信息工程研究所，北京 100093；2. 中国科学院大学网络空间安全学院，北京 100093；
3. 国家计算机网络应急技术处理协调中心，北京 100029；4. 北京航空航天大学自动化学院，北京 100191）

摘 要：为了解决遮挡场景下行人再识别的特征不对齐、错误匹配的问题，提出了一种语义引导对齐的注意力网络（SGAN）对齐行人的不同部分。SGAN 以行人的语义掩膜作为监督信息，通过全局语义引导和局部语义引导提取行人的全身和局部特征，并根据人体不同部分的可见性动态调整模型训练。在推理阶段，依据注意力模型获得局部区块的可见性，利用共享可见的人体部分的匹配策略自适应地对特征进行相似度的计算。实验结果表明，SGAN 能够容忍一定的遮挡，它的准确率不仅在全身数据集上优于大多数先进模型，在 2 个较大规模的复杂遮挡数据集 Occluded-DukeMTMC 和 P-DukeMTMC-reID 上也优于现有的行人再识别方法。

关键词：深度学习；遮挡行人再识别；注意力网络；语义引导；特征对齐

中图分类号：TN92

文献标识码：A

DOI: 10.11959/j.issn.1000−436x.2021184

Semantic guidance attention network for occluded person re-identification

REN Xuena[1,2], ZHANG Dongming[1,3], BAO Xiuguo[1,3], LI Bing[4]

1. Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China
2. School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100093, China
3. National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing 100029, China
4. School of Aeronautic Science and Engineering, Beijing University of Aeronautics and Astronautics, Beijing 100191, China

Abstract: To solve the problem of misalignment and mismatch in occluded person Re-ID, SGAN (semantic guided attention network) was proposed. In SGAN, the semantic masks of pedestrians were used as supervision to learn the global and local features through the attention modules, and the training process was dynamically adjusted according to the visibility of local regions. In the inference stage, the part-to-part matching strategy was adopted to adaptively measure visible features based on the feature visibility, which was obtained based on the learned masks from the attention modules. Experimental results show that the average accuracy of SGAN on the holistic datasets is better than most advanced models. Additionally, it is tolerant of occlusions and largely outperforms existing person Re-ID methods on two larger-scale complex occlusion datasets (Occluded-DukeMTMC and P-DukeMTMC-reID).

Keywords: deep learning, occluded person re-identification, attention network, semantic guidance, feature alignment

1 引言

行人再识别（Re-ID, re-identification）也称行人重识别，是指利用计算机视觉技术在不同监控设备采集的大规模图像或者视频中搜索目标行人的技术。随着深度学习神经网络在图像识别和计算机

收稿日期：2021−01−26；**修回日期**：2021−04−08

通信作者：张冬明，zhdm@cert.org.cn

基金项目：国家重点研发计划基金资助项目（No.2018YFB0804704）；国家自然科学基金资助项目（No.61672495, No.U1736218）

Foundation Items: The National Key Research and Development Program of China (No.2018YFB0804704), The National Natural Science Foundation of China (No.61672495, No.U1736218)

视觉领域的广泛成功应用，行人再识别取得了快速的发展，作为视频监控研究领域的关键组成部分，近几年逐渐成为研究热点，受到广泛的关注[1-2]。它可以弥补固定摄像头的视觉局限，并可与行人检测[3]、行人跟踪技术[4]相结合，应用于视频监控、智能安防、智慧城市等领域。研究基于深度学习的行人重识别的相关技术具有非常重要的理论意义和应用前景。

当前，Re-ID 方法在公开的基准数据集上的准确率基本达到了 90%以上，但是在面对复杂多变的实际场景时，模型性能会急剧下降。这是由于实际的应用场景中，行人可能受到不同程度的遮挡，遮挡物可能是物体，例如汽车、树木等，也可能是其他行人。这些遮挡导致 Re-ID 性能降低。当然，还有影响 Re-ID 的其他因素，包括拍摄姿态和视角、光照，以及采集视频的清晰度、分辨率等。当采集数据时间跨度较长时，Re-ID 还可能面临服装变化的巨大挑战。因此，研究能适应复杂场景的行人识别方法是当前的主要趋势，也是行人再识别面临的主要挑战[5-7]。本文重点研究行人识别中的遮挡问题，即遮挡行人再识别。

遮挡行人再识别用遮挡的图像作为查询对象，在不同摄像头下采集的数据中查找相同身份的行人。待查找数据中既有全身图像又有遮挡图像。如前所述，由于遮挡的随机性，全身行人再识别的有效算法在遮挡数据集上性能会严重下降，有必要研究可统一处理全身和遮挡的行人再识别框架及算法。

本文针对遮挡行人再识别中特征不对齐问题，提出了一种利用注意力对齐语义特征的算法。该算法利用注意力机制学习行人的全局特征和带有语义信息的局部特征，并根据局部特征的可见性约束特征的训练和匹配，抑制遮挡区域影响，进而实现图像对之间的共有特征的语义对齐匹配，可统一实现全身和遮挡的行人再识别。

本文主要的研究工作如下。

1) 语义引导网络。利用行人掩膜作为监督信息，以注意力的形式增加对非遮挡区域的关注，设计全局语义引导和局部语义引导结构，抑制遮挡和背景因素的影响。网络训练中，利用头部、上半身、下半身以及脚部的监督信息得到对应部分的可见性，根据可见性动态地训练模型。相比已有语义引导工作，本文将对人体语义模型嵌入网络中，构成端到端网络，仅在训练阶段使用外部语义模型结果作为监督信息，在推理阶段则不再依赖外部模型。

2) 局部特征对齐。局部语义特征分为头部、上半身、下半身和脚部，分别设计对应的注意力结构，并用对应的头部、上半身、下半身、脚部的掩膜作监督。通过局部注意力得到带有语义信息的局部特征，实现特征的语义对齐。在损失函数设计中，利用人体结构中头部、上半身、下半身以及脚部在全身中的占比分配不同的权重损失约束。

3) 基于可见性的相似度计算。利用局部语义引导学到的掩膜计算局部特征的可见性，选择待检索图像和底库图像中同时出现的特征，采用局部到局部的匹配策略得到图像间的匹配度。

实验结果表明，本文提出的方法不仅在全身数据集 Market1501[8]和 DukeMTMC-reID[4,9]上保持了较高的识别准确率，更重要的是能够有效应对遮挡问题，在复杂的遮挡数据集 Occluded-DukeMTMC[10]和 P-DukeMTMC-reID[11]上的性能优于其他先进算法。实验结果验证了本文算法可统一实现全身和遮挡的行人再识别。

2　行人再识别

2.1　全身行人再识别

基于深度学习的行人再识别方法近年取得了较大的进展。行人再识别方法大致可分为特征提取方法和基于距离度量学习的方法两类。特征提取方法核心是找到能够很好表现行人的表观特征的模型，而基于距离度量学习的方法关注的是找到有效度量行人特征相似度的准则。以下主要针对特征提取方法进行阐述。

特征提取方法的重点在于设计稳健可靠的行人图像特征表示模型，提高模型的泛化能力和稳健性，降低模型对各种影响因素的敏感度。特征提取方法又分为全局特征提取和局部特征提取。

全局特征提取利用卷积网络对整幅图像提取特征图，对特征图通过一个全局池化得到一个特征向量。利用全局特征进行行人识别通常会建模为分类和验证 2 种模型。PersonNet（person re-identification with deep convolutional neural network）[12]构建验证模型学习输入图像对的融合特征，并判断是否为同一个行人。MuDeep（multi-scale deep learning model）[13]利用分类子网络和验证子网络分别学习单幅图像的全局特征和 2 幅图像的一个

融合特征来进行类别的预测。全局特征因进行全局池化会丢失空间信息，此外，由于全局特征提取主要关注某一个身体区域，因此在着装相似场景下学习不到判别性的特征。针对全身数据的学习方式会因数据上的遮挡部分使学到的特征带有噪声，从而导致匹配错误。而实际行人数据非常复杂，单独使用全局特征不能满足性能要求，因此，局部特征提取逐渐成为当前主流的研究方法。

局部特征提取方法通过人工或者自动方法让网络关注显著的局部区域，然后提取这些区域的局部特征。常用的提取局部特征的方式主要有图像切块[14-16]、先验知识（如姿态）估计关键点定位、人体语义分割、行人前景分割等[17-18]。局部特征提取方式能从一定程度上减轻遮挡部分的影响，但简单的均匀分块的方法仍需要预先人工剪切人体区域，而姿态估计等又严重依赖外部模型的性能。

2.2 遮挡行人再识别

针对遮挡行人识别问题，当前研究工作主要集中于表征学习能力提升和不同特征匹配两方面。这些研究方法大致可以分为三类。

1) 遮挡预处理方法。这类方法首先对遮挡图像进行人工裁剪或者网络分割，去掉遮挡区域，只留下可见的部分行人区域；然后用部分行人图像进行检索。例如，DSR（deep spatial feature reconstruction）[19]和 SFR（spatial feature reconstruction）[20]处理的图像先进入全卷积网络（FCN, fully convolution network）进行分割，再利用整个库中的图像对查询图像的像素特征进行稀疏重建，在 DSR 的基础上进行提升，通过 FCN 生成多尺度特征以处理特征图的尺度问题。STNRe-ID（spatial transformer networks Re-ID）利用孪生网络输入一对图像对，图像对由同一身份的全身图像和部分图像组成，利用 STN（spatial transformer network）学习仿射变换得到仿射图像，使仿射图像逼近部分可见的图像。这类方法并不是真正的遮挡行人识别方法，对遮挡部分的处理会消耗较多的时间与人力成本。

2) 局部−全身特征匹配方法。AFPB（attention framework of person body）[11]和 T-S（teacher-student）[21]通过显著性掩膜学习遮挡图像中的显著特征与全身图像特征匹配，该方法不需要裁剪及分割的操作，比直接利用被遮挡的图像特征更容易找到相似图像。然而，局部−全局的匹配策略很明显会引起特征的不对齐问题。

3) 局部−局部特征匹配方法。Zheng 等[22]提出了一种基于字典学习的 AMC（ambiguity-sensitive matching classifier），并引入滑动窗口匹配（SWM, sliding window matching）解决全局局部匹配的问题。VPM（visibility-aware part model）[23]通过自我监督学习感知可见区域，从而避免遮挡区域的噪声影响。在测试过程中，给定待比较的 2 幅图像，VPM 首先计算它们共享区域之间的局部距离，然后得出总体距离。PVPM（pose-guided visible part matching）[24]、PGFA（pose-guided feature alignment）[10]和高阶信息[25]利用人体姿态估计得到人体的关键点信息，从而利用语义信息块进行局部特征块之间的对齐。

DSR、SFR、VPM 以及 AMC 都需要预先人工裁剪遮挡部分区域，只保留可见区域。姿态估计[10]、人体语义解析模型[26]以及显著性检测[21]方法能够直接处理遮挡图像，不需要耗时的人工裁剪，并且能够准确地定位行人关键点以及行人身体区域，实现特征对齐，但该类方法的性能对外部模型有较强的依赖性，且在测试阶段会因辅助模型引入额外的计算消耗。因此，本文提出语义引导的注意力网络（SGAN, semantic-guided attention network）来对齐不同人体部分，利用注意力学习带有语义信息的局部特征，根据局部特征块的语义信息进行对齐，并利用掩膜信息判断各个局部特征的可见性，训练过程中根据监督信息的可见程度对训练损失进行动态约束，并在测试阶段根据局部语义注意力的掩膜信息选择是否参与相似性度量。

3 语义引导对齐的遮挡行人识别

SGAN 的网络结构如图 1 所示。采用残差卷积神经网络结构 ResNet-50[27]作为主干网络，在第一个残差层 Res-Stage1 和第四个残差层 Res-Stage4 中分别增加全局语义引导和局部语义引导，并专门设计了相应的网络损失函数约束。

3.1 全局语义引导

全局语义引导和局部语义引导网络结构分别如图 2 和图 3 所示，全局语义引导包括空间注意力和通道注意力。其中，FC 代表全连接层；Conv 代表卷积操作；ReLu 和 Sigmoid 均是激活函数，实现非线性变换。

从 ResNet-50 的 Res-Stage1 得到特征图 A 后，特征图进入空间注意力层，在空间注意力层经过

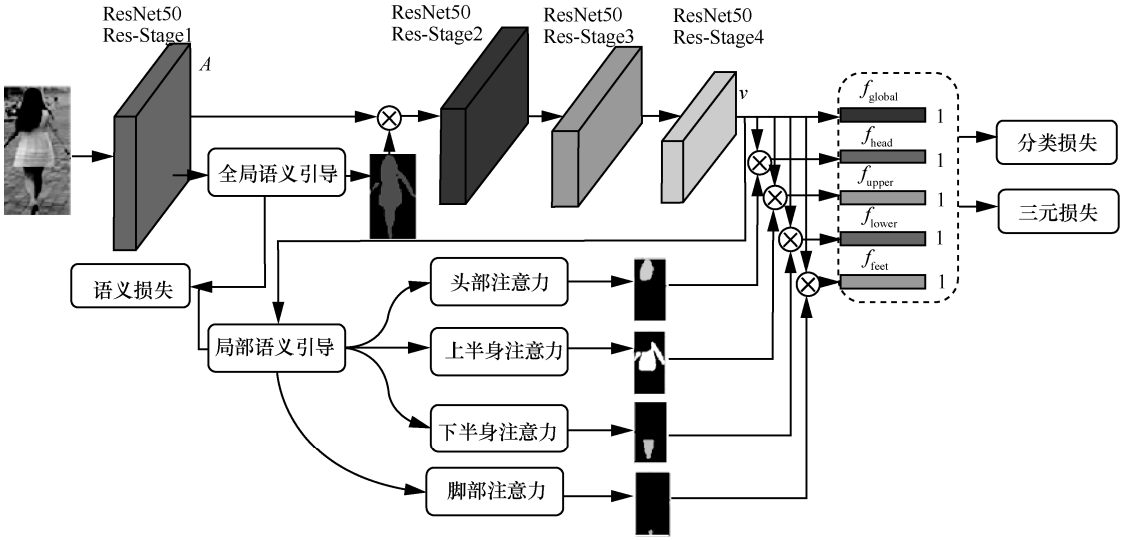

图 1　SGAN 的网络结构

一个 3×3 的卷积操作、ReLu 激活函数、1×1 的降维卷积操作得到单通道的概率图，最后经过 Sigmoid 函数获得空间注意特征图。从 ResNet-50 的 Res-Stage1 得到特征图，并与空间注意特征图做乘法得到图像中行人的前景特征图，前景特征图进一步通过通道注意力层，经过全局平均池化操作和 2 个升维降维的全连接层以及 Sigmoid 函数，最终得到过滤背景和遮挡的全局特征 f_{global}。

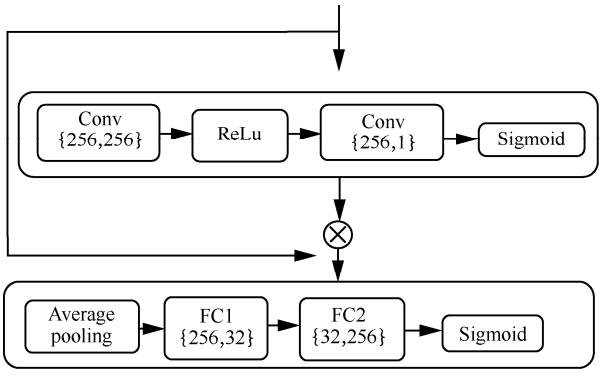

图 2　全局语义引导网络结构

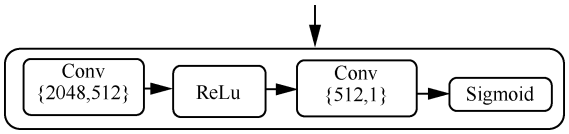

图 3　局部语义引导网络结构

3.2　局部语义引导

在 ResNet-50 的 Res-Stage4 后增加局部语义引导，在局部语义引导中有 4 个分支，分别对应头部、上半身、下半身和脚部部分。4 个局部的注意力结构相同，4 个分支的语义引导结构相同，且只有空间注意力。如图 3 所示，先对 Res-Stage4 的输出特征 V 进行降维操作，从通道 2048 变为 512，再经过 ReLu 激活函数的非线性变换，之后将 512 通道经过 1×1 的卷积得到单张通道的空间注意力特征图，得到头部、上半身、下半身和脚部的注意力特征图后，与 V 相乘分别得到局部语义特征 f_{head}、f_{upper}、f_{lower}、f_{feet}。

3.3　损失函数

三元损失[28]是 SGAN 模型训练的基本损失，SGAN 计算三元损失 L_{tri}，在此基础上，SGAN 将同时考虑全局、局部语义损失和注意力损失。下面介绍这些损失的影响因子计算方法。

3.3.1　动态因子

由于人体的头部、上身以及下身的比例不同，根据人体的头身比例特点这个先验知识来为 4 个局部特征分配不同的权重系数。

因男性与女性的身体比例差异及采集图像中行人姿态和年龄等因素的影响，本文对头部、上半身、下半身和脚部的比例分割采用不同的比例，根据人体的头身比以及人体的黄金分割比得到一个比例范围，头身比约为 1/6 到 1/8，再根据黄金比例分割参数 0.618 得到上下身比例。最终在 16×8 的监督掩膜中得到 4 个局部−全身的比例，分别为 3/16

（头部/全身）、5/16（上身/全身）、1/2（下身/全身）及 3/16（脚部/全身），比例参数用 r 表示。据此得到 4 个比例参数为

$$r_l, l \in \{\text{head}, \text{upper}, \text{lower}, \text{feet}\} \tag{1}$$

根据比例参数计算得到各个局部权重参数，通过人体语义解析模型预先得到行人数据集的语义掩膜 M，用该掩膜作为监督信息，同时用它来判断每个身体部位特征的可见性。通过计算掩膜的均值来得到 4 个局部身体特征的可见标签。每一部分的掩膜计算都会从上一特征区域的最后一行开始计算，局部特征图的大小为 16×8，行数以 0～15 为索引，则头部的掩膜均值计算对应 0～2 行，上身的掩膜均值计算对应 2～6 行，下身的掩膜均值计算对应 6～13 行，脚部的掩膜计算对应 13～15 行。先计算监督掩膜的概率均值，再与各比例参数操作得到各个局部语义部分的权重参数，按式(2)计算。

$$\text{wf}_l = \frac{\sum_{w=1}^{16}\sum_{h=1}^{8}\text{ms}(w,h)}{16\times 8 r_l} \tag{2}$$

其中，16 和 8 分别为局部监督掩膜图像的高度和宽度，$\text{ms}(w,h)$ 为监督掩膜中 (w,h) 处的像素值。

在训练阶段，从 ResNet-50 的第四个残差层得到特征 V，特征 V 进入局部语义引导部分，局部语义引导中包含 4 个分支，分别用来得到图形中行人的头部、上身、下身及脚部的概率图，用得到的 4 个身体部位的概率图分别与特征图 V 相乘，再经过平均池化层得到 4 个局部语义特征 f_{head}、f_{upper}、f_{lower}、f_{feet}，考虑遮挡因素的影响，某个局部特征可能只有极少的几个像素可见，在可见像素很少时，设定此局部特征不可见，不参加模型的训练。局部特征的可见性按式(3)计算。经实验对比，δ 设置为 0.1。

$$v_l = \begin{cases} 1, \text{wf}_l > \delta \\ 0, \text{其他} \end{cases} \tag{3}$$

3.3.2 分类损失

根据式(4)与式(5)计算局部和全局损失。

$$L_{\text{cls}\_l} = -\sum_l \left(v_l \text{wf}_l \sum_{i=1}^{C} \left((1-\xi) q_i + \frac{\xi}{C} \right) \log p_i^l \right) \tag{4}$$

$$L_{\text{cls}\_g} = -\sum_{i=1}^{C} \left((1-\xi) q_i + \frac{\xi}{C} \right) \log p_i^g \tag{5}$$

$$q_i = \begin{cases} 1, i = \text{ID} \\ 0, \text{其他} \end{cases} \tag{6}$$

其中，C 是身份类别总数，ξ 是平滑参数，p_i^l 和 p_i^g 分别是 C 维的局部向量和全局向量中的第 i 维度的预测值，ID 是行人图像的真实分类，L_{cls_l} 和 L_{cls_g} 分别对应的全局特征分类和局部特征分类。

3.3.3 注意力损失

注意力损失用来引导全局和局部语义网络的学习，用于监督的掩膜是二值掩膜，为了能够使各个语义引导部分学习到更加逼近真实掩膜的注意力概率图，采用二分类的交叉熵损失计算。局部注意力损失用 L_{s_l} 表示，同局部分类损失一样，也根据 wf_l 的值动态地约束掩膜的学习。全局注意损失用 L_{s_g} 表示。全局、局部注意力损失分别为

$$L_{\text{s}\_g} = -\frac{1}{M_g}\sum_{i=0}^{M_g}\left[b_g^i \log s_g^i + (1-b_g^i)\log(1-s_g^i) \right] \tag{7}$$

$$L_{\text{s}\_g} = -\sum_l \left(v_l \text{wf}_l \left(\frac{1}{M_l}\sum_{i=0}^{M_l}\left[b_l^i \log s_l^i + (1-b_l^i)\log(1-s_l^i) \right] \right) \right) \tag{8}$$

其中，M_g 和 M_l 分别是全局掩膜和局部掩膜，本文实验设置 M_g 的大小为 64×32，M_l 的大小为 16×8；b_g^i 和 b_l^i 分别是全局掩膜和局部掩膜在 i 位置的掩膜标签；s_g^i 和 s_l^i 分别是全局引导和局部引导在 i 位置的预测值。

这样，SGAN 的总体损失函数为

$$L_{\text{total}} = L_{\text{cls}\_g} + L_{\text{cls}\_l} + L_{\text{s}\_g} + L_{\text{s}\_l} + L_{\text{tri}} \tag{9}$$

3.4 基于可见特征行人识别度量策略

在测试阶段，根据局部语义引导部分学到的概率图来判断当前区域属于遮挡部分还是行人区域。v_l 是对应局部特征的可见性标签，如前所述，v_l=0 表示遮挡，v_l=1 表示特征可见。sim_g 表示全局特征的相似度，sim_l 表示局部特征的相似度值。sim 值越大，表明相似度越高。

$$\text{sim}=\text{sim}_g + \frac{\sum_l v_l \text{sim}_l}{1+\sum_l v_l} \tag{10}$$

本节选取 Occluded-DukeMTMC 数据集中的图像进行实验。对上述相似度计算方式进行可视化，

如图 4 所示，不可见部分对最终的相似度没有影响，可见部分对相似度的贡献取决于对应部分相似度，其权重受可见块数调节。查询目标图像与检索图像示例如图 5 所示，q 和 g 分别表示查询的目标图像和 Gallery 中的某一幅图像，其具体的匹配计算如表 1 所示。

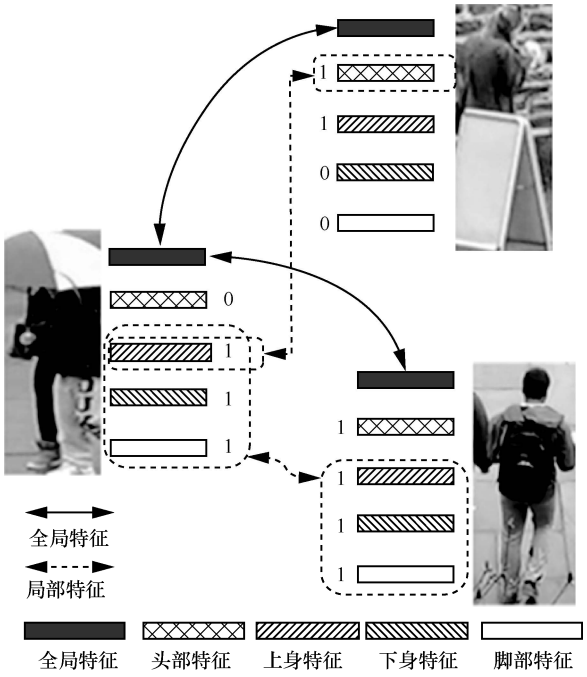

图 4　可见特征匹配策略

图 5　查询目标图像与检索图像示例

表 1　　q 与 g 的相似度匹配计算

部位	可见性		sim_l	sim_l+sin_g	$Sim(q, g)$
	q	g			
全局	1	1	1.963 3		
头部	1	1	0.025 9		
上半身	1	1	0.257 9	2.360 1	2.042 7
下半身	1	1	0.112 9		
脚部	0	1	0		

4　实验结果及分析

4.1　实验数据及评价标准

为了验证所提方法的有效性，分别在主流公开的全身数据集和遮挡数据集上进行实验验证。采用的全身数据集为 Market1501 和 DukeMTMC-reID，采用的遮挡数据集为 Occluded-DukeMTMC 和 P-DukeMTMC-reID。

Market-1501[8]包含来自 1 501 个行人的共 32 668 幅图像。这些图像来自 6 个采集设备，其中包括 5 个高分辨率摄像头和一个低分辨率摄像头。751 个行人的 12 936 幅图像被分为训练集，其余 750 个行人的 19 732 幅图像被划分为测试集。测试集中又分为待查找对象（query）和底库（gallery）。query 有 3 368 幅图像，gallery 有 19 734 幅图像。每个行人最多具有 6 幅查询图像。

DukeMTMC-reID[9]是跟踪数据集 DukeMTMC 的一个子集，舍弃了只出现在一个摄像头中的 408 个行人（即干扰项 ID），只选取出现在 2 个摄像头以上的 1 404 个行人。因此它包含来自 8 个不同摄像机的 1 404 个行人的 36 411 幅图像。702 个行人的 16 522 幅图像被分为训练集。剩余的 702 个行人的 19 889 幅图像被分为测试集，其中，query 为 2 228 幅，gallery 为 17 661 幅。

Occluded-DukeMTMC[10]包含 15 618 幅训练图像、17 661 幅 gallery 图像和 2210 幅被遮挡的 query 图像。在训练集、query、gallery 中遮挡图像所占的比例分别是 9%、100%和 10%。所有的 query 都是遮挡的图像，在 gallery 中既有全身图像又有遮挡图像。

P-DukeMTMC-reID[11]训练集 12 927 幅图像、665 个行人，query 中有 634 个行人的 2 163 幅图像，gallery 中有 9 053 幅图像。

本节采用行人识别中常用的累积匹配特征（CMC, cumulative match characteristic）曲线和平均平均精度（mAP, mean average precision）来评估所提方法。CMC 曲线中的 Rank-n 指匹配结果的前 n 幅图像的正确率。本节实验中，Rank-1、Rank-5、Rank-10 和 mAP 用来衡量算法的性能并与其他方法进行比较。所有实验结果均是在单个 q 设置下。指标及具体方法如下。

CMC。给定 Gallery 集合 G，包含 N 幅图像，分属 M 个 ID。从 Query 集合 Q 中选取给定一幅未知 ID 的图像 q，计算 q 与 G 中所有图像的相似度，对相似度结果进行排序，计算 Rank-k 如下

$$\text{Rank-}k=\begin{cases}1\text{, 前}k\text{个结果中有与}q\text{相同的ID}\\0\text{, 前}k\text{个结果中没有与}q\text{相同的ID}\end{cases} \tag{11}$$

其中，k 是一个从 1 开始增加的变量，最常用的为

1、5、10。用 Q 中所有图像的 Rank-k 值相加，再除以 Q 的总数即可得到 CMC 的值。

mAP。mAP 是 Q 中每个 q 的 AP（average precision）的平均值。AP 计算的是和 q 同一 ID 的图像在查询结果中的占比，计算式为

$$\mathrm{AP}=\frac{\left|\{\text{同ID的图片}\}\cap\{\text{查询结果}\}\right|}{\left|\{\text{查询结果}\}\right|} \tag{12}$$

4.2 实验环境及参数设置

实验平台的操作系统为 Ubuntu16.04，一块 NVIDIA 1080TI GPU，显存为 11 GB；使用深度学习框架 Pytorch1.0.1，基于 Python3.5.2 完成程序编程；使用在 ImageNet[29] 数据集上预训练的 ResNet-50 参数初始化主干网络，并去掉了全局平均池化层和全连接层。输入图像的尺寸是 256 像素×128 像素，在实验中用到了行人再识别中常用的数据增强策略，包括图像的随机水平翻转、标准化和随机擦除策略。三元损失函数中的边界超参为 0.3，训练过程中使用 Adam 优化器进行优化，训练批次是 64，每一训练批次中包括 16 个行人，每个行人 4 幅图像。初始的学习率为 3×10^{-4}，分别在 50、100 epoch 时按照 10%速度衰减，迭代次数为 200。

4.3 性能对比

为了验证模型的有效性，本节在 Market-1501、DukeMTMC-reID 和遮挡数据集 Occluded-DukeMTMC、P-DukeMTMC-reID 数据集上进行实验，表 2 和表 3 分别展示了所提 SGAN 与当前的主流方法的对比结果。

从表 2 和表 3 中可以看出，SGAN 获得了最好的表现。提取全局特征的方法会引入遮挡噪声，从而影响特征的表达；均匀分块方法没有考虑局部特征的语义性，图像中的遮挡特征会引起特征块的不对齐；SGAN 在 Occluded-DukeMTMC 数据集上的性能比 HONet[25]在 mAP 和 Rank-1 上分别提高了 2.9%和 1.6%，在 P-DukeMTMC-reID 数据集上的 mAP 比最新方法 PVPM 的结果提高了 2.2%。

表 4 展示了 SGAN 在 Market-1501 和 DukeMTMC-reID 数据集上与不同方法的性能对比结果。从表 4 可以看出，SGAN 在全身数据集上也有较好的表现，通过在行人再识别网络中引入注意力网络学习全局和局部特征，并根据注意力网络的概率图得到的语义信息进行局部特征的对齐策略有效提高了网络的精度，且明显优于其他现有方法。

表 2 Occluded-DukeMTMC 数据集上的对比结果

方法	Rank-1	Rank-5	Rank-10	mAP
Dim[30]	21.5%	36.1%	42.8%	14.4%
LOMO+XQDA[31]	8.1%	17%	22.0%	5.0%
PCB[15]	42.6%	57.1%	62.9%	33.7%
Random Erasing[32]	40.5%	59.6%	66.8%	30.0%
HACNN[33]	34.4 %	51.9%	59.4%	26.0%
DSR[19]	40.8%	58.2%	65.2%	30.4%
SFR[20]	42.3%	60.3%	67.3 %	32.0%
Part Aligned[34]	28.8%	44.6%	51.0%	20.2%
FD-GAN[35]	40.8%	—	—	—
AdverOccluded[36]	44.5%	—	—	32.2%
Part Bilinear[37]	36.9%	—	—	—
PGFA[10]	51.4%	68.6%	74.9%	37.3%
HONet[25]	55.1%	—	—	43.8%
SGAM[38]	55.1%	68.7%	74%	35.3%
SGAN	**58.0**%	**72.2**%	**78.7**%	**45.4**%

表 3 P-DukeMTMC-reID 数据集上的对比结果

方法	Rank-1	Rank-10	Rank-10	mAP
Teacher-S[21]	51.4%	50.9%	—	—
PCB[15]	79.4%	87.1%	91.0%	63.9%
IDE[9]	82.9%	89.4%	91.5%	65.9%
PVPM[24]	85.1%	91.3%	93.3%	69.9%
SGAN	**85.3**%	**92.6**%	**94.3**%	**72.1**%

4.4 性能分析

4.4.1 语义引导的有效性

为了验证 SGAN 中语义引导的有效性，以及动态训练的有效性，本节在遮挡数据集上进行了对比实验。对比实验包括 4 组，分别是行人识别基准（B）、使用全局注意力引导（G）、使用局部注意力引导（L）、同时使用全局和局部注意力引导（G+L）。实验基准使用 Resnet-50 提取图像全局特征，用身份约束和三元损失训练网络，测试时用全局特征进行度量。其余 3 组（G、L、G+L）实验在训练时各个损失带有权重约束，测试时将特征聚合。

Occluded-DukeMTMC 数据集上语义引导的对比结果如表 5 所示。从表 5 可以看出，基准实验和只有全局注意力的 Rank-1 差 1.8%，局部注意力比基准实验高 2.5%，说明利用语义引导部分能够关注图像中的可见区域，从一定程度上缓解遮挡的影响。G 和 L 的对比说明对齐的局部语义特征能够减

小匹配的误差，在只有全局特征时，会存在半身–全身的特征度量，这导致 G 比 L 的首位命中率低。G+L 中 Rank-1 达到了 58.0%，对齐的局部特征和全局特征相结合提高了模型的准确率。

表 4　Market-1501 和 DukeMTMC-reID 数据集上的对比结果

方法	Rank-1	mAP	Rank-1	mAP
BoW+kissme[8]	44.4%	20.8%	25.1%	12.2%
SVDNet[39]	82.3%	62.1%	76.7%	56.8%
PAN[40]	82.8%	63.4%	71.7%	51.5%
PAR[34]	81%	63.4%	—	—
DSR[19]	83.5%	64.2%	—	—
MultiLoss[41]	83.9%	64.4%	—	—
TripletLoss[25]	84.9%	69.1%	—	—
Adver occluded[36]	86.5%	78.3%	79.1%	62.1%
APR[42]	87%	66.9%	73.9%	55.6%
MultiScale[43]	88.9%	73.1%	79.2%	60.6%
MLFN[44]	90%	74.3%	81%	62.8%
PCB[15]	92.4%	77.3%	81.9%	65.3%
PGFA[10]	91.2%	76.8%	82.6%	65.5%
VPM[23]	93%	80.8%	83.6%	72.6%
SGAM[38]	91.4%	77.6%	83.5%	67.3%
SGAN	**93.3%**	**82.3%**	**85.5%**	**71.6%**

表 5　Occluded-DukeMTMC 数据集上语义引导的对比结果

方法	Rank-1	Rank-5	Rank-1	mAP
B	54.1%	70.1%	77.2%	43.1%
G	55.9%	72.4%	79.3%	45.9%
L	56.6%	71.7%	76.2%	42.4%
G+L	58.0%	72.2%	78.7%	45.4%

4.4.2　权重损失约束的有效性

表 6 显示了不同权重损失约束下模型在 Occluded- DukeMTMC 数据集上的性能对比结果。3 组对比实验分别是不带权重的身份损失和不带权重的注意力损失 C+A、带权重的身份损失和不带权重的注意力损失 w×C+A、带权重的身份损失和带权重的注意力损失 w×(C+A)。从表 6 可以看到，随着对分类损失和注意力损失增加约束，模型的首位命中率逐渐提高，尤其在给注意力损失增加权重约束之后，Rank-1 和 mAP 的值比不带损失约束提高了 1.4%和 0.8%。由此可以得出增加权重损失约束能够使模型更关注占比较大的可见区域，使模型以较大概率地从这些区域学习判别性的特征。

表 6　Occluded-DukeMTMC 数据集上权重损失约束的对比结果

方法	Rank-1	Rank-3	Rank-5	mAP
C+A	56.6%	73.6%	78.3%	44.7%
w×C+A	56.8%	72.6%	78.1%	43.6%
w×(C+A)	58.0%	72.2%	78.7%	45.4%

4.4.3　可视化结果

为了验证模型的全局和局部特征的学习性能，对模型学习到的全局概率图和局部语义概率图进行了可视化，实验中图像来自 Occluded-DukeMTMC 数据集。局部注意力结果和语义注意力图可视化结果分别如图 6 和图 7 所示，图 6 中数据保留 2 位小数显示。从图 6 和图 7 中可以看出，利用语义注意损失，不仅能够准确地定位到各个局部特征，还能够利用得到语义掩膜判断各个局部特征的可见性。图 8 为利用 GradCAM[45] 方法在特征图上的可视化结果，可以看出模型突出了各个身体的局部区域。图像检索结果如图 9 所示，其中显示了排序中的前 5 个图像，图像上方的 CORRECT 表示匹配正确，图 9 也说明 SGAN 在图像存在遮挡时，能够在一定程度上找到相匹配的目标图像。

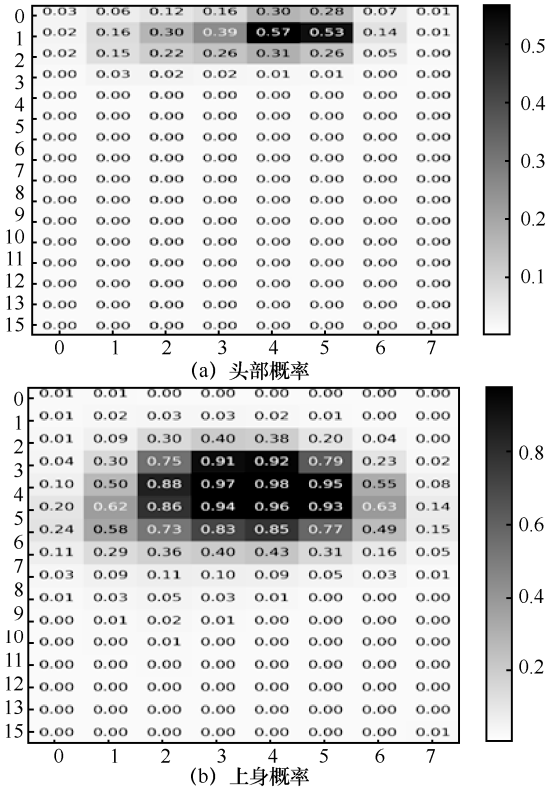

图 6　局部注意力结果

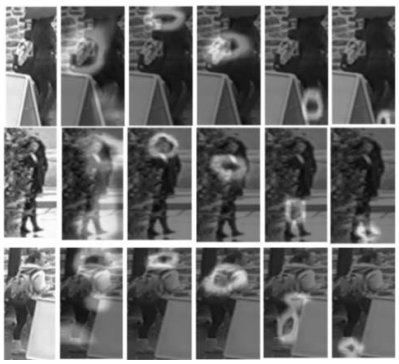

图 7　语义注意力图可视化结果

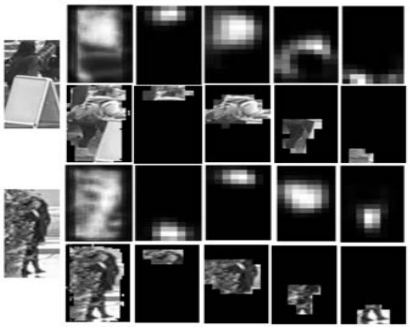

图 8　GradCAM 在特征图上的可视化结果

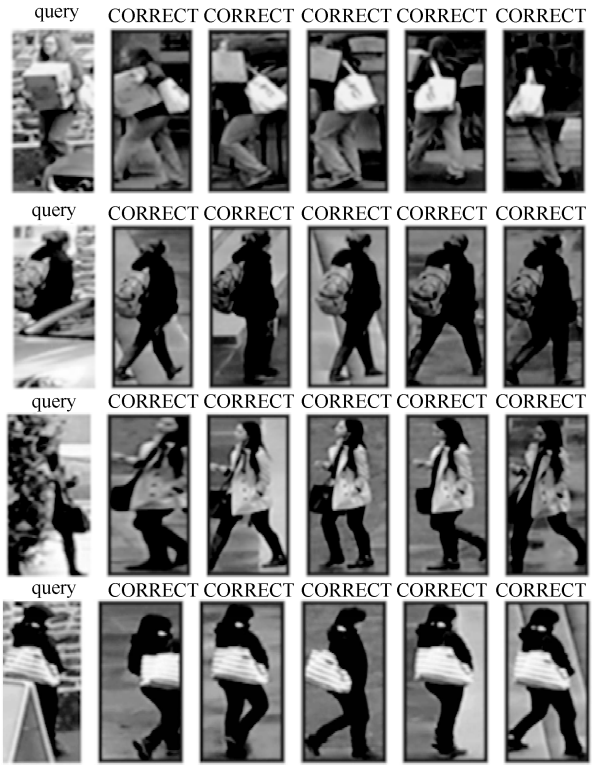

图 9　图像的检索结果

5　结束语

本文对行人再识别中的遮挡导致的特征不对齐问题进行研究。考虑遮挡的随机性以及人体结构的一般比例，提出基于语义引导进行特征对齐的遮挡行人再识别网络模型。SGAN 能够根据局部特征的可见性对模型训练和特征匹配进行约束，根据特征的可见性实现动态训练，并借助语义信息实现同语义特征块对齐及共有的可见特征的匹配计算。实现结果表明，算法获得了优异的检索性能，在复杂遮挡数据集 Occluded-DukeMTMC 和 P-DukeMTMC-reID 上算法的 Rank-1/mAP 分别达到 58.0%/45.4% 和 84.0%/71.2，优于当前的最新方法。本文研究表明，利用语义特征可有效引导模型降低遮挡区域对行人再识别的负面影响。端到端的行人再识别网络不仅能够减少对其他语义模型的依赖，还能避免因使用语义模型带来的计算消耗。

在未来的研究工作中，考虑改变网络内部结构，对遮挡进行感知，研究遮挡图像特征与全身图像特征的关联度，期望不利用附加的人体语义分割结果作为监督信息即可以实现准确遮挡行人再识别,从而提高模型的泛化能力。

参考文献：

[1] GHEISSARI N, SEBASTIAN T B, HARTLEY R. Person reidentification using spatiotemporal appearance[C]//Proceedings of 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). Piscataway: IEEE Press, 2006: 1528-1535.

[2] GRAY D, TAO H. Viewpoint invariant pedestrian recognition with an ensemble of localized features[M]. Berlin: Springer, 2008.

[3] LOWE D G. Object recognition from local scale-invariant features[C]// Proceedings of the Seventh IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 1999: 1150-1157.

[4] RISTANI E, SOLERA F, ZOU R, et al. Performance measures and a data set for multi-target, multi-camera tracking[C]//European Conference on Computer Vision. Berlin: Springer, 2016: 17-35.

[5] 罗浩, 姜伟, 范星, 等. 基于深度学习的行人重识别研究进展[J]. 自动化学报, 2019, 45(11): 2032-2049.
LUO H, JIANG W, FAN X, et al. A survey on deep learning based person Re-identification[J]. Acta Automatica Sinica, 2019, 45(11): 2032-2049.

[6] 宋婉茹, 赵晴晴, 陈昌红, 等. 行人重识别研究综述[J]. 智能系统学报, 2017, 12(6): 770-780.
SONG W R, ZHAO Q Q, CHEN C H, et al. Survey on pedestrian re-identification research[J]. CAAI Transactions on Intelligent Systems, 2017, 12(6): 770-780.

[7] ZHENG L, YANG Y, HAUPTMANN A G. Person re-identification: past, present and future[J]. arXiv Preprint, arXiv: 1610.02984, 2016.

[8] ZHENG L, SHEN L Y, TIAN L, et al. Scalable person Re-identification: a benchmark[C]//Proceedings of 2015 IEEE Interna-

tional Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2015: 1116-1124.

[9] ZHENG Z D, ZHENG L, YANG Y. Unlabeled samples generated by GAN improve the person Re-identification baseline in vitro[C]//Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2017: 3774-3782.

[10] MIAO J X, WU Y, LIU P, et al. Pose-guided feature alignment for occluded person re-identification[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2019: 542-551.

[11] ZHUO J X, CHEN Z Y, LAI J H, et al. Occluded person Re-identification[C]//Proceedings of 2018 IEEE International Conference on Multimedia and Expo (ICME). Piscataway: IEEE Press, 2018: 1-6.

[12] WU L, SHEN C, HENGEL AV. PersonNet: person re-identification with deep convolutional neural networks[J]. arXiv Preprint, arXiv: 1601.0725, 2016.

[13] QIAN X L, FU Y W, JIANG Y G, et al. Multi-scale deep learning architectures for person re-identification[C]//Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2017: 5409-5418.

[14] VARIOR R R, SHUAI B, LU J W, et al. A siamese long short-term memory architecture for human Re-identification[M]. Cham: Springer International Publishing, 2016: 135-153.

[15] SUN Y F, ZHENG L, YANG Y, et al. Beyond part models: person retrieval with refined part pooling (and a strong convolutional baseline) [C]//European Conference on Computer Vision. Berlin: Springer, 2018: 501-518.

[16] ZHANG X, LUO H, FAN X, et al. AlignedReID: surpassing human-level performance in person re-identiflcation[J]. arXiv Preprint, arXiv: 1711.08184, 2017.

[17] ZHAO H Y, TIAN M Q, SUN S Y, et al. Spindle net: person re-identification with human body region guided feature decomposition and fusion[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2017: 907-915.

[18] ZHENG L, HUANG Y J, LU H C, et al. Pose-invariant embedding for deep person re-identification[J]. IEEE Transactions on Image Processing, 2019, 28(9): 4500-4509.

[19] HE L X, LIANG J, LI H Q, et al. Deep spatial feature reconstruction for partial person re-identification: alignment-free approach[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 7073-7082.

[20] HE L, SUN Z, ZHU Y, WANG Y. Recognizing partial biometric patterns[J]. arXiv Preprint, arXiv: 1810.07399, 2018.

[21] ZHUO J, LAI J, CHEN P. A novel teacher-student learning framework for occluded person re-Identification[J]. arXiv Preprint, arXiv: 1907.03253, 2019.

[22] ZHENG W S, LI X, XIANG T, et al. Partial person Re-identification[C]//Proceedings of 2015 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2015: 4678-4686.

[23] SUN Y F, XU Q, LI Y L, et al. Perceive where to focus: learning visibility-aware part-level features for partial person re-identification[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2019: 393-402.

[24] GAO S, WANG J Y, LU H C, et al. Pose-guided visible part matching for occluded person ReID[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 11741-11749.

[25] WANG G A, YANG S, LIU H Y, et al. High-order information matters: learning relation and topology for occluded person re-identification[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 6448-6457.

[26] KALAYEH M M, BASARAN E, GÖKMEN M, et al. Human semantic parsing for person Re-identification[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 1062-1071.

[27] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2016: 770-778.

[28] HERMANS A, BEYER L, LEIBE B. In defense of the triplet loss for person re-identification[J]. arXiv Preprint, arXiv: 1703.07737, 2017.

[29] DENG J, DONG W, SOCHER R, et al. ImageNet: a large-scale hierarchical image database[C]//Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2009: 248-255.

[30] YU Q, CHANG X, SONG Y Z, et al. The devil is in the middle: exploiting mid-level representations for cross-domain instance matching[J]. arXiv Preprint, arXiv: 1711.08106, 2017.

[31] LIAO S C, HU Y, ZHU X Y, et al. Person re-identification by local maximal occurrence representation and metric learning[C]//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2015: 2197-2206.

[32] ZHONG Z, ZHENG L, KANG G L, et al. Random erasing data augmentation[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 13001-13008.

[33] LI W, ZHU X T, GONG S G. Harmonious attention network for person Re-identification[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 2285-2294.

[34] ZHAO L M, LI X, ZHUANG Y T, et al. Deeply-learned part-aligned representations for person Re-identification[C]//Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2017: 3239-3248.

[35] GE Y X, LI Z W, ZHAO H Y, et al. FD-GAN: pose-guided feature distilling GAN for robust person Re-identification[J]. arXiv Preprint, arXiv: 1810.02936, 2018.

[36] HUANG H J, LI D W, ZHANG Z, et al. Adversarially occluded samples for person Re-identification[C]//Proceedings of 2018 IEEE/CVF

Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 5098-5107.

[37] SUH Y, WANG J D, TANG S Y, et al. Part-aligned bilinear representations for person Re-identification[C]//European Conference on Computer Vision. Berlin: Springer, 2018: 418-437.

[38] YANG Q, WANG P Z, FANG Z H, et al. Focus on the visible regions: semantic-guided alignment model for occluded person Re-identification[J]. Sensors (Basel, Switzerland), 2020, 20(16): 4431.

[39] SUN Y F, ZHENG L, DENG W J, et al. SVDNet for pedestrian retrieval[C]//Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2017: 3820-3828.

[40] ZHENG Z D, ZHENG L, YANG Y. Pedestrian alignment network for large-scale person re-identification[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2019, 29(10): 3037-3045.

[41] LI W, ZHU X, GONG S. Person re-identification by deep joint learning of multi-loss classification[J]. arXiv Preprint, arXiv: 1705.04724, 2017.

[42] LIN Y T, ZHENG L, ZHENG Z D, et al. Improving person re-identification by attribute and identity learning[J]. Pattern Recognition, 2019, 95: 151-161.

[43] CHEN Y B, ZHU X T, GONG S G. Person Re-identification by deep learning multi-scale representations[C]//Proceedings of 2017 IEEE International Conference on Computer Vision Workshops (ICCVW). Piscataway: IEEE Press, 2017: 2590-2600.

[44] CHANG X B, HOSPEDALES T M, XIANG T. Multi-level factorisation net for person Re-identification[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 2109-2118.

[45] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization[J]. International Journal of Computer Vision, 2020, 128(2): 336-359.

[作者简介]

任雪娜（1989− ），女，河北石家庄人，中国科学院信息工程研究所博士生，主要研究方向为行人重识别（遮挡行人识别、变装行人识别等）。

张冬明（1977− ），男，江苏盐城人，博士，国家计算机网络应急技术处理协调中心研究员、博士生导师，主要研究方向为多媒体内容检索、模式识别、视频编码等。

包秀国（1963− ），男，江苏如皋人，博士，国家计算机网络应急技术处理协调中心教授级高级工程师、博士生导师，主要研究方向为网络与信息安全。

李冰（1990− ），男，辽宁沈阳人，北京航空航天大学博士生，主要研究方向为压缩视频行为识别。