# 基于语义融合的域内相似性分组行人重识别

寇旗旗[1]，黄绩[2]，程德强[2]，李云龙[2]，张剑英[2]
（1.中国矿业大学计算机科学与技术学院，江苏 徐州 221116；2.中国矿业大学信息与控制工程学院，江苏 徐州 221116）

**摘 要：**无监督跨域行人重识别旨在使有标签源域数据集上训练的模型适应目标域数据集。然而，基于聚类的无监督跨域行人重识别算法在网络特征学习过程中常因输入行人图片情况各异而产生噪声，从而影响聚类效果。针对这一问题，提出一种基于语义融合的域内相似性分组行人重识别网络，首先在 Baseline 网络的基础上添加语义融合层，依次从空间和通道 2 个方面对中间特征图进行相似特征的语义融合，从而提升网络的自适应感知能力。此外，通过充分利用域内相似性细粒度信息，进而提高网络对全局和局部特征的聚类精准度。通过在 DukeMTMC-ReID、Market1501 和 MSMT17 这 3 个公开数据集上进行实验，结果表明，所提算法的均值平均精度（mAP）和 Rank 识别准确率与近年无监督跨域行人重识别算法相比有显著提升。

**关键词：**无监督跨域；行人重识别；语义融合；自适应感知；细粒度信息
**中图分类号：**TP391
**文献标志码：**A
**DOI:** 10.11959/j.issn.1000−436x.2022136

# Person re-identification with intra-domain similarity grouping based on semantic fusion

KOU Qiqi[1], HUANG Ji[2], CHENG Deqiang[2], LI Yunlong[2], ZHANG Jianying[2]
1. School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China
2. School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116, China

**Abstract:** Unsupervised cross-domain person re-identification aims to adapt a model trained on a labeled source-domain dataset to a target-domain dataset. However, the cluster-based unsupervised cross-domain pedestrian re-identification algorithm often generates noise due to the different input pedestrian pictures during the network feature learning process, which affects the clustering results. To solve this problem, An intra-domain similarity grouping pedestrian re-identification network based on semantic fusion was proposed. Firstly, a semantic fusion layer was added on the basis of the Baseline network, and the semantic fusion of similar features was performed on the intermediate feature maps from the two aspects of space and channel in turn, so as to improve the adaptive perception ability of the network. In addition, by making full use of the fine-grained information of intra-domain similarity, the network's clustering accuracy of global and local features was improved. Experiments were carried out on three public datasets, DukeMTMC-ReID, Market1501, MSMT17, and the results demonstrate that the mAP and Rank recognition accuracy are significantly improved compared with recent unsupervised cross-domain person re-identification algorithms.

**Keywords:** unsupervised cross domain, person re-identification, semantic fusion, adaptive perception, fine-grained information

## 0 引言

行人重识别[1-2]任务的目标是在同一区域内的多个摄像机视角中识别并匹配具有相同身份的人，它在智能监控系统中发挥着重要作用。该任务可以分为有监督和无监督 2 种情况，近年来，有监督重

**收稿日期：**2022−03−21；**修回日期：**2022−06−15
**通信作者：**程德强，chengdq@cumt.edu.cn
**基金项目：**中央高校基本科研业务费专项资金资助项目（No.2020QN49）
**Foundation Item:** The Fundamental Research Funds for the Central Universities (No.2020QN49)

识别任务所取得的优异成果给学术界留下了深刻印象，但由于训练数据集包含标签，不仅标注成本巨大，而且在实际测试时不具备实时获取目标域标签的能力，导致监督行人重识别难以满足实际应用的需求[3]。此时，无监督训练的优势便体现出来，利用有标签的源域数据集训练出具有较强泛化性的网络，应用于无行人标签的目标域，这类网络称为无监督跨域行人重识别网络。

在网络跨域训练过程中，为了解决标签问题，通常采用聚类的方式为行人分配伪标签，节省了人工标注的成本。深度卷积神经网络通过堆叠卷积层和池化层来学习判别特征，由于输入行人图片情况各异，如行人身体错位和区域比例不一致等，导致识别的准确率受影响。其中，身体错位一般有 2 种情况：1) 人在行走时被相机抓拍导致姿态不同；2) 由于检测不完善，导致同一行人在不同图像中的身体部位出现区域比例不一致问题。在网络对特征向量进行聚类时，上述问题产生的噪声会直接影响聚类结果的准确性。

此外，在域自适应过程中不同数据域相机风格或背景风格等存在差异性，这种差异性对网络的泛化能力是一种巨大的考验。为了缩小这种差异，目前有 2 种主流方法：1) 通过增强数据集或网络重新生成数据集的方式，加大训练样本的数量来提高网络识别性能[4-5]；2) 基于生成对抗网络（GAN, generative adversarial network）将图像外观从源域转换到目标域，从而增加 2 个域的相关性[6-7]。上述针对数据集操作的方法均是对源域和目标域之间相关性的考虑，目标域内训练样本中存在的相似性并未被进一步挖掘，且在网络学习过程中增加了额外计算成本。

针对图像身体错位等因素导致聚类结果不准确的问题，本文提出一种简洁高效的基于语义融合的域内相似性分组网络。本文的主要贡献如下。

1) 本文网络在 Baseline 网络的基础上创新性地添加了两层语义融合层，实现对网络中间特征图的细化处理，增强卷积神经网络提取特征的辨识度，其中，本文提出的语义融合层包含空间语义融合（SSF, spatial semantic fusion）和通道语义融合（CSF, channel semantic fusion）2 个模块。

2) 在不增加额外计算成本的前提下，本文利用域内行人的细粒度相似性特征，将网络的输出特征图水平分割为两部分，通过聚类的方法根据全局和局部各自的域内相似性对行人进行分类，使同一行人被分配多个伪标签，构成新的数据集。被分配相同伪标签的不同行人图片具有许多相似性，通过新的数据集对预训练模型进行微调来迭代挖掘更精确的行人分类信息。

3) 与近年会议中提出的算法相比较，本文算法在 DukeMTMC-ReID、Market1501 和 MSMT17 这 3 个公共数据集上的跨域识别率得到显著提升，算法的直接效果通过热图以及检索排序等方式进行展示。

## 1 相关工作

### 1.1 跨域行人重识别

最近，众多学者密切关注跨域行人重识别算法，利用在源域中训练的重识别模型以提高对未标记目标域行人的识别性能，跨域行人重识别也称作无监督域自适应行人重识别，它解决了不同域间差异性的挑战。但是，由于源域训练的模型对目标域中特征变化很敏感，在使用预训练模型适应目标域时必须考虑到图像的变化，当前无监督域自适应行人重识别的解决方案可以分为三类：图像风格迁移、中间特征对齐和基于聚类的方法[8]。

在图像风格迁移方法中使用基于生成对抗网络[9]是当下流行的方法。ECN（exemplar-camera-neighborhood）[10]利用迁移学习并使用示例记忆最小化目标不变性来学习不变特征；多视图生成网络 CR-GAN（context rendering GAN）[6]着眼于背景风格，通过掩盖目标域图像中的行人以保留背景杂波，叠加源域中行人和目标域背景作为输入图像来训练模型。但是，GAN 的训练过程复杂，而且会引入额外的计算成本，因此不适用于实际场景。

中间特征对齐方法旨在减少域间特征和图像级别的差距，假设源域数据集和目标域数据集共享一个共同的中间特征空间，该共同中间特征可以用于跨域推断人员身份。D-MMD 损失（dissimilarity-based maximum mean discrepancy loss）[11]通过使用小批量来关闭成对距离，实现特征对齐；基于补丁的无监督学习（PAUL, patch-based unsupervised learning）[12]框架假设如果两幅图像相似，那么图像间存在相似的局部补丁；PAUL[12]并不学习图像全局级别特征，而是为行人识别提供局部细节级别特征。

基于聚类的方法通常根据聚类结果生成硬伪标签或软伪标签，然后根据带有伪标签的图像训练模型和交替迭代这 2 个步骤使模型达到最优。深度软多标签参考学习模型 MAR[13]根据特征相似性和分类概率

之间的差异挖掘潜在的成对关系，然后使用对比损失加强挖掘的成对关系；UDAP（unsupervised domain adaptive person re-identification）[4]计算重排序的距离后对目标图像进行聚类，然后根据聚类结果生成伪标签；SAL（self-supervised agent learning）[14]算法通过利用一组代理作为桥梁来减少源域和目标域之间的差异。

上述 3 种域自适应行人重识别方法在训练时通过缩小源域和目标域之间的差距从而提高模型的泛化能力，然而忽略了目标域内同一行人自身存在一定的相似性。利用这一特性，本文对目标域行人特征进行上下分块，聚焦于行人图像上下部分的非显著性特征，用聚类的方法将两部分特征进行聚类，为行人共分配 3 种伪标签。

### 1.2　建模尺度变化

针对公共数据集内存在的图像尺寸和人物比例不一致的问题，近年已有研究增强对尺寸和比例变化的特征表示能力。传统方法一般采用尺寸不变的特征变换，如 SIFT（scale invariant feature transform）[15]和 ORB（oriented FAST and rotated BRIEF）[16]；对于卷积神经网络，通过图像对称、尺度变换和旋转等操作对数据进行转换。然而，此类方法采用固定尺寸的卷积核进行操作，导致其对于未知的转换任务存在局限性。此外，一些其他方法自适应地从数据域中学习空间转换：STN（spatial transformer network）[17]通过全局参数变换来扭曲特征图；DCN（deformable convolutional network）[18]用偏移量增加了卷积中的采样位置，并通过端到端的反向传播来学习偏移量。

上述方法均通过对网络进行大数据量的训练来得到图像变换参数，这对于数据量有限的行人识别任务来说并不合适。本文提出的空间语义融合模块计算空间语义相似度，对相同身体部位信息进行聚集，无须进行参数训练。而且，在语义融合层中的通道语义融合模块通过建模计算通道之间存在的相关性，显著增强了特征的表示能力。

## 2　基本原理

参照现有的大多数跨域识别网络在源域数据集上对模型进行预训练的方式，本文利用在 ImageNet[19] 上预训练好的 ResNet50[20] 作为 Baseline 网络。如图 1 所示，在 Baseline 网络 $layer_2$ 和 $layer_3$ 后分别添加语义融合层（虚线框内 2 个深灰色层）作为主干网，为中间特征图融合更多语义信息。将原网络最后的全连接（FC，fully connected）层替换为两层维度分别为 2 048 和源域身份数的全连接层。将网络输出的特征图 $\boldsymbol{F}$ 水平切分为上下两块 $\boldsymbol{F}_{\text{up}}$ 和 $\boldsymbol{F}_{\text{dn}}$，由此可以获取更多的细粒度特征。分别对特征图 $\boldsymbol{F}$、$\boldsymbol{F}_{\text{up}}$ 和 $\boldsymbol{F}_{\text{dn}}$ 进行全局平均池化（GAP, global average pooling）操作得到特征向量。然后将不同行人图像的特征向量分组并分配伪标签。通过最小化每组伪标签的三元组损失 $L_{\text{tri}}$ 来迭代更新模型。

### 2.1　语义融合层

语义融合层依次对空间和通道信息进行融合。空间语义融合模块根据输入行人图像的姿态和尺度自适应地确定感受野。给定来自卷积神经网络的中间特征图，利用相似特征和相邻特征之间的高相关性特点，自适应地定位各种姿势和不同比例的身体部位，以此来更新特征图。将更新后的特征图经过批量归一化（CBN, batch normalization）层与原特征图构成残差结构，再将结果进行通道语义融合。通道语义融合模块是通道之间的相关语义融合，实现小规模视觉线索的保留。图 2 为语义融合

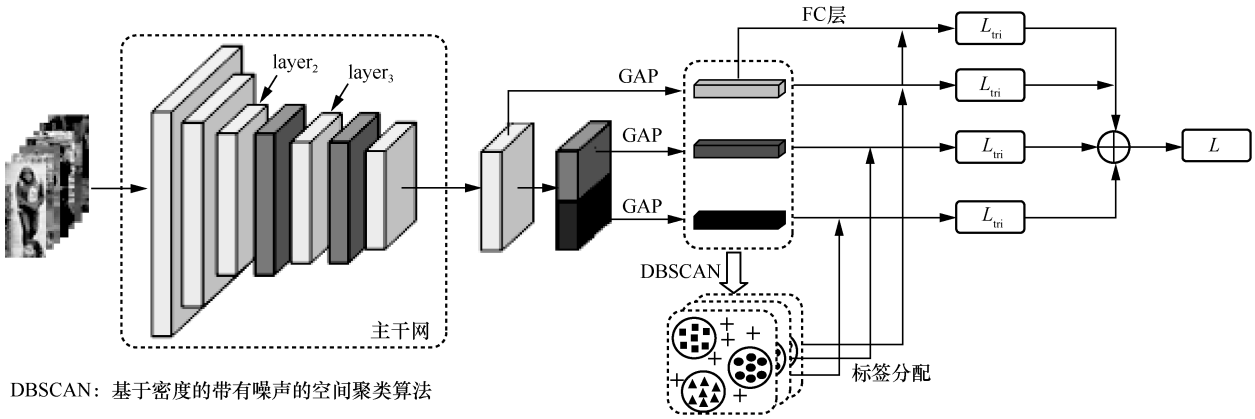



图 1　整体网络结构

层的网络结构，残差结构可以使融合层保持良好的性能。

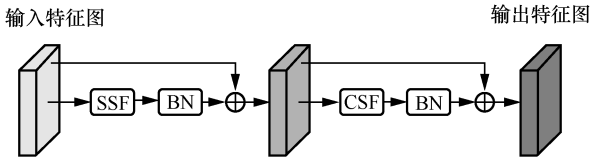



图 2 语义融合层的网络结构

### 2.1.1 空间语义融合模块

受限于卷积神经网络的固定网络结构，卷积层在固定位置对特征图进行采样，池化层以固定比例降低空间分辨率。由于特征图感受野一般为矩形，导致感受野对行人不同姿态适应性较差。此外，固定大小的感受野对于不同尺寸的身体部位进行编码是不合适的。为了解决这个问题，本文对中间特征图进行空间语义融合，通过建模空间特征的相互依赖关系，自适应地确定每个特征的感受野，从而提高特征对身体姿势和比例变化的稳健性。

空间语义融合模块如图 3 所示。假设给定一个特征图 $\boldsymbol{F}\in R^{C\times H\times W}$，其中 $C$、$H$ 和 $W$ 分别表示通道数、特征图高度和宽度。首先，将 $\boldsymbol{F}$ 重塑为 $\boldsymbol{F}\in R^{C\times M}$，其中 $M$ 为空间特征的数量（$M=H\times W$）；然后，从特征图的外观关系和位置关系两方面对空间特征进行依赖性建模，生成语义关系图 $\boldsymbol{S}$；最后，融合特征图 $\boldsymbol{F}$ 和语义关系图 $\boldsymbol{S}$，生成新的融合特征图。

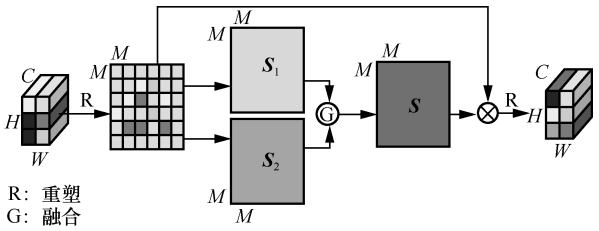



图 3 空间语义融合模块

对于外观关系，通过测量输入特征图中任意两位置之间的外观相似性来生成外观关系映射图。Du 等[21]提到在相邻空间位置的局部特征具有重叠的感受野，所以它们之间有较高的相关性。因此涉及相邻位置的感受野可以获得更精细的外观。假设 $f_i,f_j\in R^C$ 表示特征图 $\boldsymbol{F}$ 中第 $i$ 个和第 $j$ 个空间位置的特征，分别选取 $i$ 和 $j$ 位置周围大小为 $E\times E$ 的感受野，然后通过累加相应位置特征之间的点积，使用 SoftMax 函数对 $\boldsymbol{F}$ 中的所有空间位置进行归一化处理得到外观相似性，计算式为

$$\left(\boldsymbol{S}_1^E\right)_{ij}=\frac{\exp\left(\sum\limits_{e=1}^{E\times E}\left(\boldsymbol{p}_{i,e}^{\mathrm{T}}\boldsymbol{p}_{j,e}\right)\right)}{\sum\limits_{t=1}^{H\times W}\exp\left(\sum\limits_{e=1}^{E\times E}\left(\boldsymbol{p}_{i,e}^{\mathrm{T}}\boldsymbol{p}_{t,e}\right)\right)} \tag{1}$$

其中，$\boldsymbol{p}_{i,e}$ 和 $\boldsymbol{p}_{j,e}$ 分别表示感受野大小为 $e$ 的 $i$ 和 $j$ 位置上的特征，$\boldsymbol{S}_1^E$ 表示感受野大小为 $E$ 对应的外观关系图。

根据式(2)融合不同尺寸 $E$ 的感受野，得到对身体部位更稳健的关系图。SoftMax 函数可以抑制不同部位较小的相似度，通过式(2)可以得到外观关系图 $\boldsymbol{S}_1$。

$$\boldsymbol{S}_1=\mathrm{SoftMax}\left(G\left(S_1^1,\cdots,S_1^Q\right)\right) \tag{2}$$

其中，$G$ 为具有元素乘积的融合函数，$Q$ 为不同尺度感受野的数量。

对于位置关系，行人图像对应于相同的身体部位特征在空间上相近，通过二维高斯函数可以计算空间特征 $f_i$ 和 $f_j$ 之间的位置关系，即

$$l_{ij}=\frac{1}{2\pi\sigma_1\sigma_2}\exp\left[-\frac{1}{2}\left(\frac{(x_j-x_i)^2}{\sigma_1^2}+\frac{(y_j-y_i)^2}{\sigma_2^2}\right)\right] \tag{3}$$

其中，$(x_i,y_i)$ 和 $(x_j,y_j)$ 分别为 $f_i$ 和 $f_j$ 的位置坐标，$(\sigma_1,\sigma_2)$ 为二维高斯函数的标准差。通过式(4)规范化 $l_{ij}$，使其关系值之和为 1，记位置关系图为 $\boldsymbol{S}_2$。

$$(\boldsymbol{S}_2)_{ij}=\frac{l_{ij}}{\sum\limits_{t=1}^{H\times W}l_{it}} \tag{4}$$

最后，根据式(5)将外观关系图和位置关系图进行融合，得到空间语义关系图 $\boldsymbol{S}$。

$$\boldsymbol{S}=\mathrm{SoftMax}\left(G(\boldsymbol{S}_1,\boldsymbol{S}_2)\right) \tag{5}$$

为了在原特征图内融入空间特征，通过两者相乘的方式得到融合特征图 $\boldsymbol{F}_s$，计算式为

$$\boldsymbol{F}_s=\boldsymbol{F}\boldsymbol{S}^{\mathrm{T}} \tag{6}$$

### 2.1.2 通道语义融合模块

通常，卷积神经网络经过下采样处理后会丢失很多细节信息，然而这些细粒度信息对于行人的区分往往起到重要的作用，比如在困难样本对中，通过利用衣服纹理或背包等细节信息，可以区分 2 个不同的身份。根据 Zhang 等[22]提到的大多数高级特征的通道图对特定部分会表现出不同反应，融合不同通道中的相似特征，也可以增强行人独有的特征。

通道语义融合模块如图 4 所示。同空间语义融合一样，重塑特征图为 $\boldsymbol{F}\in R^{C\times M}$，将得到的 $\boldsymbol{F}$ 和自

身转置矩阵 $\boldsymbol{F}^{\mathrm{T}}$ 相乘，并将结果进行归一化处理得到通道关系图 $\boldsymbol{C} \in R^{C\times C}$，计算式为

$$\boldsymbol{C}=\frac{\exp\left(\boldsymbol{f}_m^{\mathrm{T}}\boldsymbol{f}_n\right)}{\sum_{l=1}^{C}\exp\left(\boldsymbol{f}_m^{\mathrm{T}}\boldsymbol{f}_l\right)} \tag{7}$$

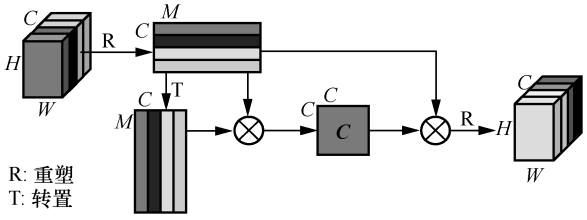



图 4　通道语义融合模块

其中，$\boldsymbol{f}_m$ 和 $\boldsymbol{f}_n$ 分别表示 $\boldsymbol{F}$ 的第 $m$ 和第 $n$ 通道中的特征。通过式(8)将通道关系图和原特征图进行融合得到新的融合特征图 $\boldsymbol{F}_c$。

$$\boldsymbol{F}_c=\boldsymbol{C}\boldsymbol{F} \tag{8}$$

### 2.2　细粒度信息的密度聚类

受到 Wang 等[23]提出的监督训练分割方法的启发，即从细粒度中可以提取出更多有用的信息。考虑到目标数据集中行人特征从全局到局部存在潜在的相似性，本文利用密度聚类方法[24]对全局和局部特征进行聚类，结合这两部分信息能够获得更稳健和有辨识度的行人特征表示。网络中语义融合层很大程度降低了可能因数据集产生的聚类噪音。

将目标域数据集中每个未标记图像 $x^i$ 输入主干网中提取特征，经过语义融合层，网络输出特征图表示为 $\boldsymbol{F}^i \in R^{C\times H\times W}$，然后将 $\boldsymbol{F}^i$ 水平分切分为两块，分别表示为 $\boldsymbol{F}_{\mathrm{up}}^i \in R^{C\times\frac{H}{2}\times W}$ 和 $\boldsymbol{F}_{\mathrm{dn}}^i \in R^{C\times\frac{H}{2}\times W}$。通过如此切分可以获取行人图像上下两部分包含信息的细粒度特征，有助于后期行人的分类。对于全局特征图 $\boldsymbol{F}^i$ 以及分块特征图 $\boldsymbol{F}_{\mathrm{up}}^i$ 和 $\boldsymbol{F}_{\mathrm{dn}}^i$ 经过全局平均池化得到特征向量 $\boldsymbol{f}^i$、$\boldsymbol{f}_{\mathrm{up}}^i$ 和 $\boldsymbol{f}_{\mathrm{dn}}^i$。每张图像经过如此操作，则有

$$\begin{cases} f=\{\boldsymbol{f}^1,\cdots,\boldsymbol{f}^N\} \\ f_{\mathrm{up}}=\{\boldsymbol{f}_{\mathrm{up}}^1,\cdots,\boldsymbol{f}_{\mathrm{up}}^N\} \\ f_{\mathrm{dn}}=\{\boldsymbol{f}_{\mathrm{dn}}^1,\cdots,\boldsymbol{f}_{\mathrm{dn}}^N\} \end{cases} \tag{9}$$

对于式(9)中的每组特征向量，利用密度聚类算法得到相应的伪标签组，即每个身份根据它所属的组分配一个伪标签。经过主干网后，每张图像 $x^i$ 对应 3 个伪标签，分别表示为 $y^i$、$y_{\mathrm{up}}^i$ 和 $y_{\mathrm{dn}}^i$。因此，可以基于 3 个特征向量分组结果组成一个有标签的数据集 $X$，如式(10)所示。此外，如图 1 所示，特征向量 $\boldsymbol{f}^i$ 通过一个维度为 2 048 的全连接层，旨在获取一个全局嵌入向量 $\boldsymbol{f}_{\mathrm{fc}}^i$，其伪标签与特征向量 $\boldsymbol{f}^i$ 共享。

$$X=\{x^i:\{y^i,y_{\mathrm{up}}^i,y_{\mathrm{dn}}^i\};1\leqslant i\leqslant N\} \tag{10}$$

### 2.3　损失函数

为了学习到更具判别力的特征，本文在预训练网络损失函数上联合使用难样本挖掘的三元组损失和 SoftMax 交叉熵损失。为每个小批量随机采样 $P$ 个身份和 $K$ 个实例，以满足难样本三元组损失的要求。三元组损失函数为

$$L_{\mathrm{tri}}=\sum_{i=1}^{P}\sum_{a=1}^{K}\left[\alpha+\max_{p=1,\cdots,K}\left\|x_a^{(i)}-x_p^{(j)}\right\|_2-\min_{\substack{n=1,\cdots,K\\ j=1,\cdots,P\\ j\neq i}}\left\|x_a^{(i)}-x_n^{(j)}\right\|_2\right]_+ \tag{11}$$

其中，$x_a^{(i)}$ 是锚点，$x_p^{(j)}$ 是与 $x_a^{(i)}$ 具有相同标签的正样本，$x_n^{(j)}$ 是标签与 $x_a^{(i)}$ 不同的负样本；$\alpha$ 是三元组损失的边缘超参数，用来控制样本对间的相对距离。难样本三元组损失使 $x_p^{(j)}$ 的欧氏距离比 $x_n^{(j)}$ 更接近锚点 $x_a^{(i)}$，即具有相同标签的样本之间将比具有不同标签的样本更接近。使用难样本三元组损失来区分难样本可促进更好的聚类，提高模型性能。

对于 Baseline 网络的训练，利用 SoftMax 交叉熵损失提高网络判别学习能力，其计算式为

$$L_{\mathrm{SoftMax}}=-\sum_{i=1}^{P}\sum_{a=1}^{K}\log\frac{\exp\left(\boldsymbol{W}_{y_{a,i}}^{\mathrm{T}}x_{a,i}\right)}{\sum_{k=1}^{H}\exp\left(\boldsymbol{W}_k^{\mathrm{T}}x_{a,i}\right)} \tag{12}$$

其中，$y_{a,i}$ 为第 $i$ 个身份的 $K$ 张图像中第 $a$ 张图像的真实标签，$H$ 为身份的数量。通过式(13)将 2 种损失函数进行组合，从而实现对预训练网络的更新。

$$L_{\mathrm{baseline}}=L_{\mathrm{tri}}+L_{\mathrm{SoftMax}} \tag{13}$$

对于域迁移网络的训练，目标域图片输入网络后，将聚类生成的伪标签作为监督信息，使用三元组损失对预训练模型进行跨域自适应微调。损失函数包含全局、上分块、下分块、全局嵌入 4 个部分，计算式为

$$L = L_{\text{tri}}(f, y) + L_{\text{tri}}(f_{\text{fc}}, y) + L_{\text{tri}}(f_{\text{up}}, y_{\text{up}}) + L_{\text{tri}}(f_{\text{dn}}, y_{\text{dn}}) \tag{14}$$

## 3 实验及结果分析

### 3.1 实验数据集

实验主要在 3 个行人数据集上对网络进行评估，包括 Market1501[25]、DukeMTMC-ReID[26]和 MSMT17[27]。

Market1501[25]数据集图像由 6 台相机捕捉，共包含身份 1501 个，总图像数量达到 32 668 张。其中，训练集身份有 751 个，图像有 12 936 张；query 图像共有 3 368 张，身份有 750 个；gallery 图像共有 15 913 张；身份有 751 个。

DukeMTMC-ReID[26]数据集是由 8 台相机捕捉的包含 1 812 个不同行人的重识别公开数据集，其中有 1 404 个身份同时出现在 2 台及以上的相机中，其余 408 个身份用作干扰项。数据集包含训练集图像共有 16 522 张，身份有 702 个；query 图像共有 2 228 张，身份有 702 个；gallery 图像共有 17 661 张，身份有 1 110 个。

MSMT17[27]数据集是一个接近真实场景的大型数据集，由 15 个相机捕捉图像共有 126 441 张，身份有 4 101 个。其中训练集图像有 30 248 张，身份有 1 041 个；query 图像有 11 659 张，身份 3 060 个；gallery 图像共有 82 161 张，身份有 3 060 个。

### 3.2 实验细节和评估指标

如第 1 节所述，首先对 Baseline 用源域数据集进行训练，采用 Zhong 等[32]使用的方法进行训练。将输入图片的大小调整为 256×128，采用随机裁剪、翻转和随机擦除对数据进行增强；为满足难样本三元组损失的要求，将每个 mini-batch 用随机选择的 $P=16$ 个身份进行采样，并从训练集中为每个身份随机采样 $K=8$ 张图片，得到 mini-batch 为 128 张，将三元组损失的边缘参数 $\alpha$ 设置为 0.5。空间语义融合模块中感受野的数量 $Q$ 设置为 3（如式(2)）。由于 ResNet[20]不同阶段特征图空间大小不同，因此本文采用不同的标准差（如式(3)），添加到 $\text{layer}_2$ 后的语义融合层 $\sigma_1$ 和 $\sigma_2$ 设置为 10 和 20，添加到 $\text{layer}_3$ 后的语义融合层 $\sigma_1$ 和 $\sigma_2$ 设置为 5 和 10。在训练中使用权重衰减为 0.000 5 的 Adam[33]优化器来优化 70 个 epoch 的参数。初始学习率设置为 $6\times10^{-5}$，在 7 个 epoch 后将学习率调整为 $1.8\times10^{-5}$，再经过 7 个 epoch 学习率调整为 $1.8\times10^{-6}$，一直训练到结束。

### 3.3 与先进算法的比较

在 3 个公共数据集上，将本文算法与近年顶级会议文章所提出的算法进行比较。将行人重识别任务通用的累积匹配特性中的 Rank 识别准确率（R-1、R-5、R-10）和均值平均精度（mAP, mean average precision）作为评价指标，评价模型在数据集上的性能。比较结果如表 1 和表 2 所示，所有数据均不经过重排序处理。

**表 1** **不同算法在 DukeMTMC-ReID 和 Market1501 的实验结果**

| 算法 | DukeMTMC-ReID→Market1501 | | | | Market1501→DukeMTMC-ReID | | | |
|---|---|---|---|---|---|---|---|---|
| | mAP | R-1 | R-5 | R-10 | mAP | R-1 | R-5 | R-10 |
| ARN | 39.4% | 70.3% | 80.4% | 86.3% | 33.4% | 60.2% | 73.9% | 79.5% |
| PAUL | 53.2% | 72.0% | 82.7% | 86.0% | 40.1% | 68.5% | 82.4% | 87.4% |
| CDS | 39.9% | 71.6% | 81.2% | 84.7% | 42.7% | 67.2% | 75.9% | 79.4% |
| CR-GAN | 54.0% | 77.7% | 89.7% | 92.7% | 48.6% | 68.9% | 80.2% | 84.7% |
| PDA-Net | 47.6% | 75.2% | 86.3% | 90.2% | 45.1% | 63.2% | 77.0% | 82.5% |
| UCDA | 34.5% | 64.3% | – | – | 36.7% | 55.4% | – | – |
| MAR | 48.0% | 67.1% | 79.8% | 84.2% | 40.0% | 67.7% | 81.9% | 87.3% |
| ECN | 43.0% | 75.1% | 87.6% | 91.6% | 40.4% | 63.3% | 75.8% | 80.4% |
| UDAP | 53.7% | 74.7% | 86.9% | 90.3% | 49.0% | 68.4% | 80.1% | 83.5% |
| D-MMD | 48.8% | 70.6% | 87.0% | 91.5% | 46.0% | 63.5% | 78.8% | 83.9% |
| NSSA | 47.9% | 76.2% | 88.7% | 92.4% | 45.5% | 65.5% | 77.9% | 81.3% |
| SAL | 38.7% | 65.3% | 79.7% | 84.6% | 48.5% | 67.6% | 80.9% | 84.7% |
| DCJ | 51.4% | 74.5% | 83.8% | 87.0% | 50.9% | 68.3% | 79.4% | 83.6% |
| 本文算法 | 56.3% | 78.6% | 88.7% | 92.0% | 52.4% | 71.7% | 79.9% | 83.0% |

**表 2**　　**不同算法在 MSMT17 的实验结果**

| 算法 | DukeMTMC-ReID→MSMT17 | | | | Market1501→MSMT17 | | | |
|---|---|---|---|---|---|---|---|---|
| | mAP | R-1 | R-5 | R-10 | mAP | R-1 | R-5 | R-10 |
| ECN[10] | 10.2 | 30.2 | 41.5 | 46.8 | 8.5 | 25.3 | 36.3 | 42.1 |
| NSSA[30] | 12.1 | 32.8 | 43.1 | 48.3 | 10.6 | 28.9 | 38.2 | 43.3 |
| D-MMD[11] | 15.3 | 34.4 | 51.1 | 58.5 | 13.5 | 29.1 | 46.3 | 54.1 |
| ECN+GPP[34] | 16.0 | 42.5 | 55.9 | 61.5 | 15.2 | 40.4 | 53.1 | 58.7 |
| MMCL[35] | 16.2 | 43.6 | 54.3 | 58.9 | 15.1 | 40.8 | 51.8 | 56.7 |
| 本文算法 | 17.4 | 45.3 | 56.2 | 62.8 | 15.9 | 42.2 | 53.0 | 57.8 |

不同算法在 MSMT17 的实验结果如表 1 所示，包括 8 种通过聚类形成伪标签的算法 UDAP[4]、MAR[13]、ECN[10]、CDS[29]、UCDA[5]、SAL[14]、DCJ[31]和 NSSA[30]；2 种通过域风格迁移的算法 CR-GAN[6]和 PDA-Net[7]；3 种特征对齐算法 ARN[28]、D-MMD[11]和 PAUL[12]。其中，CR-GAN[6]在 DukeMTMC-ReID 泛化到 Market1501 的 mAP 和 R-1 表现最好，本文算法在网络复杂度上远低于 CR-GAN[6]，而且 mAP 提高 2.3%，R-1 提高 0.9%。在数据集 Market1501 泛化到 DukeMTMC-ReID 的结果中，本文算法表现更好，和上述算法中表现最好的 DCJ[31]相比 mAP 提高了 1.5%，R-1 提高了 3.4%。

表 2 为 DukeMTMC-ReID 和 Market1501 分别泛化到 MSMT17 的实验结果。MSMT17 数据集包含的身份更多且摄像头视角更多，数据集包含较多存在身体错位和遮挡等问题的图片，更接近现实场景，难度较大。与表 2 中性能最优的 MMCL[35]算法相比，本文算法在 DukeMTMC→MSMT17 上 mAP 提高 1.2%，R-1 提高 1.7%；在 Market1501→MSMT17 上 mAP 提高 0.8%，R-1 提高 1.4%。

### 3.4　消融实验

本节首先将模型在 DukeMTMC-ReID 数据集上进行预训练，然后在 Market1501 数据集上进行消融研究，最后通过实验分别验证语义融合层中各部分和特征细粒度分块的有效性。

在添加的语义融合层内，空间语义融合模块中感受野尺寸 $E$（如式(1)）的选择对识别准确率有较大影响。如表 3 所示，不同尺寸 $E$ 的感受野较 Baseline 识别准确率均有所提高，但当 $E$ 进一步增大到 5 时，准确率开始下降。感受野的不断增大会忽略一些关键身份信息。本文在式(2)中对不同感受野对应的关系图进行融合时，选取感受野数量 $Q=3$ 得到最优的实验结果。

**表 3**　**不同感受野尺寸 $E$ 的感受野对实验结果的影响**

| $E$ | DukeMTMC-ReID→Market1501 | | | |
|---|---|---|---|---|
| | mAP | R-1 | R-5 | R-10 |
| $E$=1 | 53.7 | 75.9 | 86.2 | 90.3 |
| $E$=2 | 55.8 | 78.1 | 87.9 | 91.9 |
| $E$=3 | 56.3 | 78.6 | 88.7 | 92.0 |
| $E$=5 | 55.7 | 76.2 | 84.4 | 89.6 |

对于融合函数 $G$ 的选取，本文实验将逐元素求最大值、累加以及相乘 3 种函数作比较，实验数据如表 4 所示。在 $Q=3$ 的情况下，融合函数对经过尺度分别为 1、2、3 的感受野所获得的外观相似图进行融合，从表 4 中可知，对应位置逐元素求最大值、累加和相乘的融合函数较 Baseline 网络的识别准确率均有所提升，其中逐元素相乘的融合函数对结果提升最为显著。

**表 4**　　**融合函数 $G$ 对实验结果的影响**

| $G$ | DukeMTMC-ReID→Market1501 | | | |
|---|---|---|---|---|
| | mAP | R-1 | R-5 | R-10 |
| 最大值 | 53.9 | 77.0 | 87.5 | 89.8 |
| 累加 | 54.4 | 76.9 | 88.0 | 90.4 |
| 相乘 | 56.3 | 78.6 | 88.7 | 92.0 |

对于网络的整体结构，本节分别对语义融合层中空间语义融合和通道语义融合模块进行消融实验，实验结果如表 5 所示。通过分析，Baseline 网络分别添加空间语义融合和通道语义融合模块对识别准确率均有所提升。将二者按先空间后通道的方式串联到一起，组合成语义融合层添加到 Baseline 网络中，对识别准确率的提升最大：mAP 提高 4%，R-1 提高 3.1%。由此可见，添加语义融合层可以获取更多有效的行人特征信息，从而提高识别准确率。

**表 5　不同语义模块对实验结果的影响**

| 语义模块 | DukeMTMC-ReID→Market1501 | | | |
|---|---|---|---|---|
| | mAP | R-1 | R-5 | R-10 |
| Baseline（不添加语义模块） | 52.3 | 75.5 | 85.7 | 88.9 |
| Baseline+空间语义融合 | 54.1 | 77.5 | 87.3 | 90.3 |
| Baseline+通道语义融合 | 53.8 | 77.1 | 87.4 | 90.3 |
| 本文算法 | 56.3 | 78.6 | 88.7 | 92.0 |

对于网络输出特征图，本节在水平分块的数目上进行了消融实验。通过表 6 可知，将网络输出特征图分为上下两部分能得到最佳识别准确率。通过分析可知，当不进行分块时，特征图丢失了有用的细粒度信息；当分块较多时，由于数据集图像内存在一些身体错位和被遮挡的图像，导致在经过密度聚类时会产生较多噪声信息和较差的相似性挖掘以及匹配。因此，本文对网络结构的设计时将分块数确定为 2。

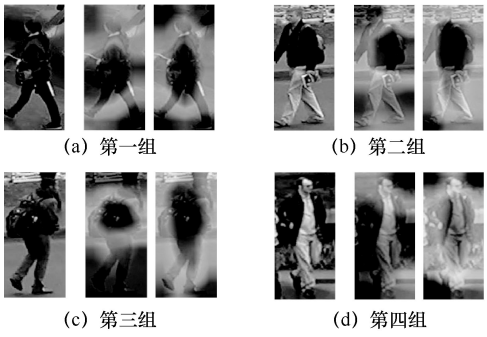
**表 6　不同分块数对实验结果的影响**

| 分块情况 | DukeMTMC-ReID→Market1501 | | | |
|---|---|---|---|---|
| | mAP | R-1 | R-5 | R-10 |
| 不分块 | 51.3 | 74.2 | 83.6 | 87.7 |
| 分 2 块 | 56.3 | 78.6 | 88.7 | 92.0 |
| 分 3 块 | 49.2 | 70.0 | 81.8 | 84.5 |
| 分 4 块 | 40.5 | 67.2 | 79.2 | 83.7 |

### 3.5　可视化分析

为了更直观地体现网络在 Baseline 上的改进，本节使用 DukeMTMC-ReID 数据集进行预训练，使用 Market1501 数据集进行训练和测试，使用热图[36]和检索排序对实验结果进行可视化分析。

热图共有 4 组图片，如图 5 所示。每组图中，第一张图像为 Market1501 数据集行人图片，第二张为经过 Baseline 网络的热图，第三张为经过本文网络的热图。从图 5 中可以看出，Baseline 网络由于固定感受野，所以只关注行人的局部信息，当图像整体色调相近时（如图 5(a)所示），Baseline 网络对行人的关注会被背景所干扰，本文方法将不同尺寸的感受野进行融合，实现了更关注行人主体的效果；当背景较为复杂时（如图 5(d)所示），Baseline 网络的关注完全偏离了人物，而本文的改进网络表现依旧稳定。



(a) 第一组　(b) 第二组　(c) 第三组　(d) 第四组

图 5　热图

图 6 分别展示了 Baseline 网络和本文网络在 Market1501 数据集上识别实例的检索排序结果。每张行人图像上方的“√”和“×”分别表示查询结果的正确与否。可以看到经过本文网络的实验结果在 R-1、R-5 上的识别准确率都较高且稳定。其中，第二组行人的衣着相似难以辨认，Baseline 网络在第二位置识别错误的行人图像在本文网络的识别结果排序中排第八位，且本文网络未出现其他识别错误图像。由此可见，在面对特征相似的行人图像时，本文网络依旧可以得到很好的识别效果。

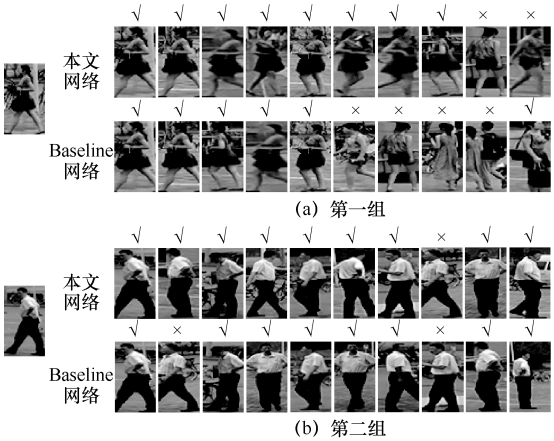



图 6　检索排序结果

## 4　结束语

本文提出了一种基于语义融合的域内相似性分组网络。语义融合层对于行人图片自适应生成不同尺度的感受野，增强了空间特征之间的相互依赖关系，通过融合通道信息进一步提高了网络的表示能力。实验结果表明，相比于未添加语义融合层前的网络，本文网络的 mAP 提高 4.0%。此外，本文提出的网络采用分块的方式对目标域内细粒度相似性信息进行挖掘，得到更精确的行人分类信息。实验数据表明，分块聚类相比于未进行分块处理的

网络 mAP 提高 5.0%。为了进一步增强网络在现实环境中的泛化性，在后续的工作中本文将采用不同光照和尘雾环境的数据集对网络进行训练。对于行人被遮挡的情况，本文会为网络添加行人遮挡模块使网络具备一定的抗遮挡能力。

## 参考文献：


[1] LI J H, CHENG D Q, LIU R H, et al. Unsupervised person re-identification based on measurement axis[J]. IEEE Signal Processing Letters, 2021, 28: 379-383.

[2] ZHAO K, CHENG D Q, KOU Q Q, et al. Sequences consistency feature learning for video-based person re-identification[J]. Electronics Letters, 2022, 58(4): 142-144.

[3] 任雪娜, 张冬明, 包秀国, 等. 语义引导的遮挡行人再识别注意力网络[J]. 通信学报, 2021, 42(10): 106-116.
REN X N, ZHANG D M, BAO X G, et al. Semantic guidance attention network for occluded person re-identification[J]. Journal on Communications, 2021, 42(10): 106-116.

[4] SONG L C, WANG C, ZHANG L F, et al. Unsupervised domain adaptive re-identification: theory and practice[J]. Pattern Recognition, 2020, 102: 107173.

[5] QI L, WANG L, HUO J, et al. A novel unsupervised camera-aware domain adaptation framework for person re-identification[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2019: 8079-8088.

[6] CHEN Y B, ZHU X T, GONG S G. Instance-guided context rendering for cross-domain person re-identification[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2019: 232-242.

[7] LI Y J, LIN C S, LIN Y B, et al. Cross-dataset person re-identification via unsupervised pose disentanglement and adaptation[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2019: 7918-7928.

[8] LIN X T, REN P Z, YEH C H, et al. Unsupervised person re-identification: a systematic survey of challenges and solutions[J]. arXiv Preprint, arXiv: 2109.06057, 2021.

[9] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.

[10] ZHONG Z, ZHENG L, LUO Z M, et al. Invariance matters: exemplar memory for domain adaptive person re-identification[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2019: 598-607.

[11] MEKHAZNI D, BHUIYAN A, EKLADIOUS G, et al. Unsupervised domain adaptation in the dissimilarity space for person re-identification[C]//Computer Vision - ECCV 2020. Cham: Springer International Publishing, 2020: 159-174.

[12] YANG Q Z, YU H X, WU A C, et al. Patch-based discriminative feature learning for unsupervised person re-identification[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2019: 3628-3637.

[13] YU H X, ZHENG W S, WU A C, et al. Unsupervised person re-identification by soft multilabel learning[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2019: 2143-2152.

[14] JIANG K, ZHANG T, ZHANG Y, et al. Self-supervised agent learning for unsupervised cross-domain person re-identification[J]. IEEE Transactions on Image Processing: a Publication of the IEEE Signal Processing Society, 2020, 29: 8549-8560.

[15] ZHAO R, OUYANG W L, WANG X G. Unsupervised salience learning for person re-identification[C]//Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2013: 3586-3593.

[16] RUBLEE E, RABAUD V, KONOLIGE K, et al. ORB: an efficient alternative to SIFT or SURF[C]//Proceedings of 2011 International Conference on Computer Vision. Piscataway: IEEE Press, 2011: 2564-2571.

[17] JADERBERG M, SIMONYAN K, ZISSERMAN A, et al. Spatial transformer networks[C]//Proceedings of the 28th International Conference on Neural Information Processing. Cambridge: MIT Press, 2015: 2017-2025.

[18] DAI J F, QI H Z, XIONG Y W, et al. Deformable convolutional networks[C]//Proceedings of 2017 IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2017: 764-773.

[19] 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.

[20] 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. Piscataway: IEEE Press, 2016: 770-778.

[21] DU Y, YUAN C F, LI B, et al. Interaction-aware spatio-temporal pyramid attention networks for action classification[C]//Computer Vision - ECCV 2018. Cham: Springer International Publishing, 2018: 388-404.

[22] ZHANG S S, YANG J, SCHIELE B. Occluded pedestrian detection through guided attention in CNNs[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 6995-7003.

[23] WANG G S, YUAN Y F, CHEN X, et al. Learning discriminative features with multiple granularities for person re-identification[C]//Proceedings of the 26th ACM International Conference On Multimedia. New York: ACM Press, 2018: 274-282.

[24] ESTER M, KRIEGEL H P, SANDER J, et al. A density-based algo-

rithm for discovering clusters in large spatial databases with noise[C]//Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. Palo Alto: AAAI Press, 1996: 226-231.

[25] ZHENG L, SHEN L Y, TIAN L, et al. Scalable person re-identification: a benchmark[C]//Proceedings of 2015 IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2015: 1116-1124.

[26] 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. Piscataway: IEEE Press, 2017: 3774-3782.

[27] WEI L H, ZHANG S L, GAO W, et al. Person transfer GAN to bridge domain GAP for person re-identification[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 79-88.

[28] LI Y J, YANG F E, LIU Y C, et al. Adaptation and re-identification network: an unsupervised deep transfer learning approach to person re-identification[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Piscataway: IEEE Press, 2018: 285-2856.

[29] WU J L, LIAO S C, LEI Z, et al. Clustering and dynamic sampling based unsupervised domain adaptation for person re-identification[C]// Proceedings of 2019 IEEE International Conference on Multimedia and Expo. Piscataway: IEEE Press, 2019: 886-891.

[30] ZHAO Y R, LU H T. Neighbor similarity and soft-label adaptation for unsupervised cross-dataset person re-identification[J]. Neurocomputing, 2020, 388: 246-254.

[31] GE Y, LIU L, ZHANG H X. A three-stage learning approach to cross-domain person re-identification[J]. Applied Soft Computing, 2021, 112: 107793.

[32] ZHONG Z, ZHENG L, ZHENG Z D, et al. Camera style adaptation for person re-identification[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 5157-5166.

[33] KINGMA D P, BA J. Adam: a method for stochastic optimization[J]. arXiv Preprint, arXiv:1412.6980, 2014.

[34] WANG D K, ZHANG S L. Unsupervised person re-identification via multi-label classification[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 10978-10987.

[35] ZHONG Z, ZHENG L, LUO Z M, et al. Learning to adapt invariance in memory for person re-identification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(8): 2723-2738.

[36] 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]

**寇旗旗**（1988− ），男，河南襄城人，博士，中国矿业大学讲师，主要研究方向为智能信息处理、计算机视觉、模式识别。

**黄绩**（1995−），男，山西忻州人，中国矿业大学硕士生，主要研究方向为无监督行人重识别。

**程德强**（1979− ），男，河南洛阳人，博士，中国矿业大学教授、博士生导师，主要研究方向为机器视觉与模式识别、图像处理与视频编码、图像智能检测与信息处理。

**李云龙**（1997− ），男，河南开封人，中国矿业大学硕士生，主要研究方向为无监督行人重识别。

**张剑英**（1964− ），女，江苏徐州人，博士，中国矿业大学教授，主要研究方向为信息处理、电磁场理论及应用等。