文章编号：2095-2163(2022)11-0018-08 中图分类号：TP391 文献标志码：A

无监督学习三元组用于视频行人重识别研究

蔡江琳，韩 华，王春媛，潘欣宇，芮行江
（上海工程技术大学 电子电气工程学院，上海 201620）

摘 要：在智能交通中，对于目前产生的海量视频通过人工来标记行人图像不切实际，使无监督学习得到更多的关注。针对在无监督学习数据中缺少详细的身份信息，无法知晓目标图像对应的正负样本问题，提出一种无监督学习三元组用于视频行人重识别研究的方法。该方法从无标签的数据集中挖掘三元组、即目标图像，与目标图像身份相同的轨迹和与目标图像身份不同的轨迹。首先根据单相机内轨迹的时空一致性，即构成轨迹的任意帧图像具有相同的身份，将行人轨迹特征表示成图像特征均值后，通过计算 *rank* - 1 轨迹作为判断三元组的条件，用于设计特殊的三元组损失函数。并根据特征距离大小分配样本权重，着重学习困难样本，使模型动态调整正、负样本对之间的距离，加速模型的收敛速率，降低过拟合风险。然后通过计算跨相机 *rank* - 1，合并高度关联的轨迹作为跨相机三元组的锚样本用于损失计算。最后联合单相机和跨相机的损失评估模型。经过实验证明，该方法在 PRID2011、iLIDS-VID 和 MARS 上的结果都表明了该模型的有效性和可靠性。

关键词：无监督学习；行人轨迹；关联排序；时空一致性；三元组损失

Unsupervised learning triplets for video-based pedestrian reidentification

CAI Jianglin, HAN Hua, WANG Chunyuan, PAN Xinyu, RUI Xingjiang
(School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 201620, China)

【Abstract】 In view of the massive videos of the intelligent transportation system, it is impractical to manually label pedestrian images, making unsupervised learning get more attention. Aiming at the lack of detailed identity information in the unsupervised learning data and the inability to know the positive and negative samples corresponding to the target image, an unsupervised learning triplet method is proposed for video pedestrian re-identification research. From an unlabeled dataset, the method mines triples, namely target images, trajectories with the same identity as the target image and trajectories with different identities from the target image .First, according to the spatio-temporal consistency of the trajectory within a single camera, that is, any frame images that constitute the trajectory have the same identity. After the pedestrian trajectory feature is expressed as the feature mean of the image, the *rank* - 1 trajectory is calculated as the condition for judging the triplet and is used for designing a special triplet loss function. Based on this, the paper assigns the sample weight according to the feature distance, focuses on learning difficult samples, makes the model dynamically adjust the distance between positive and negative sample pairs, accelerates the convergence rate of the model, and reduces the risk of overfitting. Then, by computing the cross-camera *rank* - 1, the highly correlated trajectories are merged as anchor samples for the cross-camera triples for loss computation. Finally, joint single-camera and cross-camera losses are evaluated. Experiments show that the results of this method on PRID2011, iLIDS-VID and MARS all demonstrate the validity and reliability of the model.

【Key words】 unsupervised learning; pedestrian track; association ranking; temporal and spatial consistency; triplet loss

0 引 言

行人重识别的目的是在配备多台摄像机、且视野不交叉的环境中找到具有相同身份的目标行人。当目标行人穿过某台摄像机视野时，可以在另一台摄像机下找到相同身份的人。当前的行人重识别多是基于 2 类：基于图像的行人重识别[1-7]和基于视频的行人重识别[8-12]。从传统的特征提取方法和度量学习方法、到利用卷积神经网络训练模型，基于图像的行人重识别模型已经取得了很高的识别准确度。但在实际的监控视频中，由于行人的许多不确定因素，例如光照、遮挡、姿态变化等，导致监控跟踪失败，基于图像的二维特征很难解决这些问题。而不同于图像重识别的是，基于视频的行人重识别的研究对象是行人轨迹，包含了行人更多的时空信息，连续的帧图像之间有着密切联系。当前基于视频的

基金项目：国家自然科学基金(61305014)；上海市自然科学基金(22ZR1426200)；上海市教育委员会和上海市教育发展基金会“晨光计划”(13CG60)。

作者简介：蔡江琳(2001-)，女，本科生，主要研究方向：目标识别与跟踪、图像处理；韩 华(1983-)，女，博士，教授，主要研究方向：目标识别与跟踪、行人重识别、智能计算等；王春媛(1983-)，女，博士，副教授，主要研究方向：多源信息协同处理、模式识别、机器学习等。

通讯作者：韩 华 Email：2070967@mail.dhu.edu.cn

收稿日期：2022-01-06

行人重识别技术已取得有效的成果。例如，Times Shift Dynamic Warping（TSDTW）[13] 模型通过对每个行人的时空动态信息进行编码来生成一种潜在的特征表示，解决不准确和不完整序列的选择和数据匹配问题。又如一种顶推度量学习模型[9]，是通过优化最小类内的变化来提高 top rank 中行人重识别的准确度。再如，采用一种视频排序函数[14] 方法，在排序的同时可以从含噪或者不完整的视频序列中选择可靠的时空特征。上述的视频重识别技术多是采用有监督的学习方法，但是在实际的场景中，往往不具有可测量性和实际性。为此，基于半监督[1,12] 和无监督[13-19] 的学习方法开始得到更多的关注。

当前由于无监督学习存在的一些固有性质，导致无监督模型的性能比有监督模型差。而事实上，这些基于视频研究的无监督模型不能有效地利用深度卷积神经网络[20]（deep Convolutional Neural Networks）强大的特征学习能力，获取具有表达性的特征和具有判别力的匹配模型。主要是因为无标签数据集中并不具备有效的监督信息供模型训练。在基于深度神经网络的行人重识别中，常用三元组损失函数作为度量模型损失的方法。而对于无监督学习则需要模型自主挖掘三元组用于损失计算。本文中，基于无监督学习挖掘三元组方案的主要内容有：

（1）单相机内的时空一致性，每条行人轨迹中的图像都属于同一个 *ID*，目的是利用构成轨迹的图像更新轨迹特征。

（2）从无标签数据集中挖掘三元组，设计一种自适应加权的条件、即三元组损失函数，动态调整正负样本对之间的距离，提高模型性能。

1　方法

1.1　单相机关联学习

单相机内关联学习的目的是为了学习具有判别力的单相机轨迹特征。基于单相机内的时空一致性如图 1 所示。

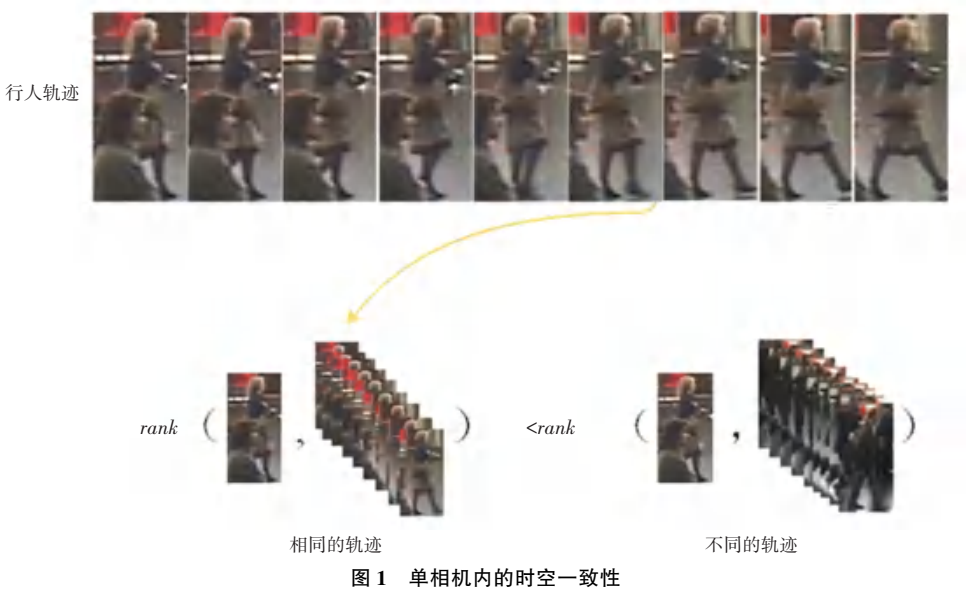


图 1　单相机内的时空一致性

Fig. 1　The spatio-temporal consistency with a single camera

研究中，将图 1 中的行人轨迹定义为源轨迹，假设摄像机 k 下有 N_k 条小片段轨迹，则所有锚样本轨迹的特征集合为 $\{a_{k,i}\}_{i=1}^{N_k}$，构成源轨迹中的任一帧图像特征表示为 $x_{k,p}$。而单相机内的时空一致性则意味着来自同一条轨迹的大多数图像都表示同一个行人，因此构成源轨迹中的任一帧图像与该轨迹间的特征距离会比该帧图像与其它轨迹间的特征距离小。

1.1.1　轨迹特征表示方法

由图 1 可知，这里的行人轨迹是由连续的图像帧构成。每条轨迹包含了同一行人的多张连续图像，为此可以提取行人丰富的时空信息，使学习到的行人特征更加具有表达力。当前的许多轨迹特征处理方案多是利用卷积神经网络中的时间池化层，例如最大池化层[21]（max-pooling）、或是平均池化层[10]（mean-pooling），将小片段轨迹表示成一种序列级的特征。但是，这种方法在网络学习的过程中需要大量的计算成本，因为每个小批量学习迭代中都要用到前馈（feed-forward）轨迹中的所有图像，会造成时间浪费。为此，在模型训练过程中，将小片段

轨迹表示成 $a_{k,i}$，并采用指数滑动平均的方法(EMA)，通过构成该轨迹中的任意帧图像 $x_{k,p}$ 来更新轨迹特征。进而推得的数学公式可写为：

$$a_{k,i}^{t+1} \leftarrow a_{k,i}^{t} - \tau(l_2(a_{k,i}^{t}) - l_2(x_{k,p}^{t})), \text{ if } i = p \quad (1)$$

其中，$i = p$ 表示更新轨迹的图像是源轨迹中的任意帧图像；t 表示小批量样本集训练迭代的次数；τ 是向 EMA 提供的衰减率参数，通常用来控制模型的更新速度；$a_{k,i}^{t}$ 相当于一个影子变量，其初值可表示为构成 $a_{k,i}$ 这条轨迹的所有图像帧的特征均值，最终目的是获取更新后轨迹特征值 $a_{k,i}^{t+1}$。

由于轨迹特征 $a_{k,i}$ 和图像特征 $x_{k,p}$ 之间存在尺度和单位的差异，研究中采用 l_2 对其进行归一化，例如，$\| l_2(\cdot) \| = 0.5$。采用指数滑动平均的算法来更新轨迹，究其原因就在于对滑动窗口中的值求平均时，前面的值都是呈指数衰减的，导致原来的值对更新后的值产生的影响减少，而最近的值权重更大，从而使滑动均值只与最近的迭代有关系。当 $a_{k,i}$ 初始化为所有图像的特征均值并根据式(1)进行迭代更新时，单相机内的锚样本会伴随着模型学习的过程持续学习来表示每条轨迹。

1.1.2 关联排序

在模型学习的过程中，逐渐更新摄像机 k 内的 N_k 轨迹特征。由式(1)获取所有锚样本轨迹集合为 $\{a_{k,i}\}_{i=1}^{N_k}$。要搜索的目标行人图像为 $x_{k,p}$，为了找到和目标图像 $x_{k,p}$ 最近邻的轨迹特征，将目标图像与摄像机 k 内的所有轨迹进行关联，计算彼此间的相似程度，并进行排序，得到一个排序列表，再找到与目标图像距离最近的轨迹特征。

本节将使用标准的 l_2 度量方法，对图像特征和轨迹特征进行标准化后将计算两者间的特征距离。计算目标图像与所有锚样本轨迹间的特征距离，这里需用到的数学公式可写为：

$$\{D_{p,i} \mid D_{p,i} = \| l_2(x_{k,p}) - l_2(a_{k,i}) \|_2, \ i \in N_k\} \quad (2)$$

其中，$\{D_{p,i}\}_{i=1}^{N_k}$ 表示 $x_{k,p}$ 与摄像机 k 中所有轨迹 $\{a_{k,i}\}_{i=1}^{N_k}$ 间的特征距离。将所有的特征距离进行排序，找到距离 $x_{k,p}$ 最小的轨迹特征，为此推得如下的数学表述形式：

$$\{D_{p,i}\}_{i=1}^{N_k} \xrightarrow{\text{在摄像机 } k \text{ 内排序}} D_{p,t} = \min_{i \in [1, N_k]} D_{p,i} \quad (3)$$

1.1.3 挖掘三元组和损失函数设计

在本节中，采用一种特殊的三元组损失函数来评估模型性能。在训练过程中起到一种类似顶推(top-push)的作用。单相机关联学习过程如图 2 所示。图 2 中，使 $rank-1$ 的轨迹 $a_{k,t}$ 能够对应于目标图像所在的轨迹 $a_{k,p}$，即 $p = t$。

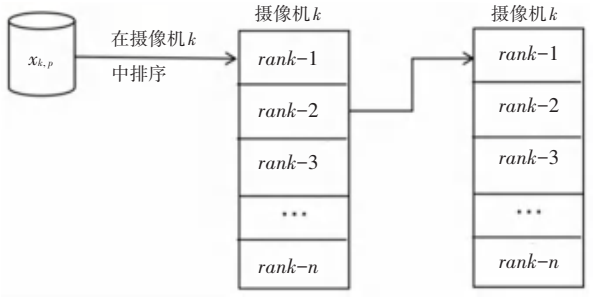


图 2 单相机关联学习过程

Fig. 2 The process of intra-camera association learning

传统的三元组损失函数是由 FaceNet[23-24] 提出，包括：锚样本 x_a、即要寻找的目标样本，与目标样本具有相同身份的正样本 x_p，与目标样本不具有相同身份的负样本 x_n，此处的数学公式具体如下：

$$L_3 = \left[\sum_{x_a, x_p} D_{a,p} - \sum_{x_a, x_n} D_{a,n} + m \right]_+ \quad (4)$$

其中，$[\cdot]_+ = \max(0, \cdot)$；$D_{a,p}$ 表示目标样本与正样本之间的特征距离；$D_{a,n}$ 表示目标样本与负样本之间的特征距离；m 是给定的阈值，可以使目标样本与正样本之间的最大距离远小于目标样本与负样本之间的最小距离。

为了在训练过程中学习更好的特征，充分挖掘各个样本对之间潜在的关联性、从而提取更加鲜明的行人特征，为此引入一种自适应加权的方法，将损失函数中的各个样本对距离加上相应的权重来训练模型，图 3 给出的就是样本权重描述。则一般加权三元组的数学计算公式见如下：

$$L_{3\_weighted} = \left[\sum_{x_p \in P} \omega_p D_{a,p} - \sum_{x_n \in N} \omega_n D_{a,n} + m \right]_+ \quad (5)$$

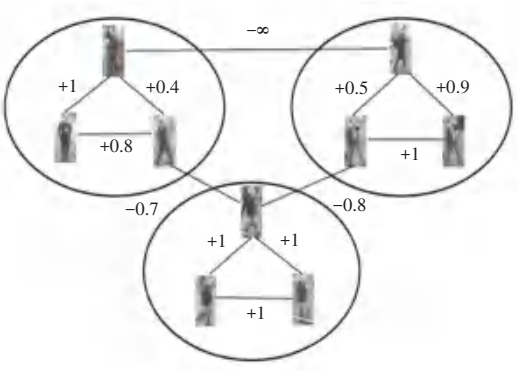
其中，$x_p \in P$ 表示正样本集，$x_n \in N$ 表示负样本集。

而由 Hermans 等人提出的困难三元组损失[21]，仅考虑正负样本时，对应的权重可以写成：

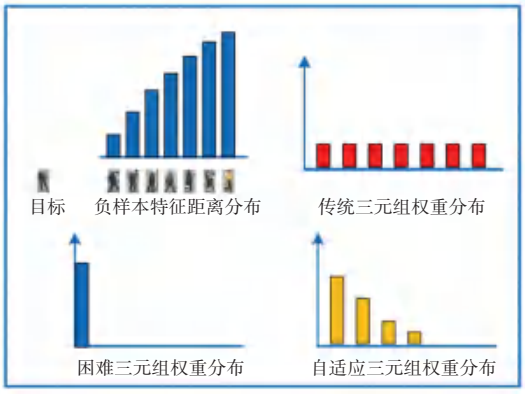
$$\begin{cases} \omega_p = [x_p == \arg\max_{x \in P} d(x_a, x)]_+ \\ \omega_n = [x_n == \arg\min_{x \in N} d(x_a, x)]_+ \end{cases} \quad (6)$$

其中，最困难正样本是指视觉上看不是同一个人、但实际是相同身份的行人，则两者之间的特征距离会最大。最困难负样本是指视觉上看是同一个人、但实际上不是相同身份的行人，则两者之间的特征距离将会最小。这种方法可以有效避免在训练过程中由于简单样本的影响使训练陷入了较坏的局部最小值。而传统的权重统一的三元组损失在模型训

练过程中对异常值较鲁棒，为此拟结合这 2 种损失的优越性，来设计本节的三元组损失函数。



（a）自适应权重



（b）权重分布

图 3　样本权重

Fig. 3　Illustration of weights

由于该模型是基于无监督的一种端到端的训练模式，因此没有预先标记的成对行人标签。为此要先找到对应的三元组，从而设计损失函数。此后的设计过程可做研究阐释如下。

由式(2)可以得到，摄像机 k 内所有锚样本轨迹 $\{a_{k,i}\}_{i=1}^{N_k}$ 与目标图像 $x_{k,p}$ 之间的特征距离 $\{D_{p,i}\}_{i=1}^{N_k}$。为了确定对应的正负样本，利用式(3)找到 $rank-1$ 的轨迹 $a_{k,t}$，并且在理想状况下可认为 $a_{k,t}$ 对应 $x_{k,p}$ 所在的轨迹 $a_{k,p}$。那么如果 $p=t$，则 $rank-1$ 轨迹 $a_{k,t}$ 就是轨迹 $a_{k,p}$，对应目标图像 $x_{k,p}$ 为正样本集；如果 $p\neq t$，则 $rank-1$ 轨迹 $a_{k,t}$ 不是轨迹 $a_{k,p}$，对应目标图像 $x_{k,p}$ 为负样本集。基于此，单相机内三元组损失可进一步剖析阐述如下。

(1)当 $p\neq t$ 时。损失函数为：

$$L_{I\_weighted}=\Big[\sum_{x_p\in a_{k,p}}\omega_p d(x_{k,p},x_p)-\sum_{x_t\in a_{k,t}}\omega_n d(x_{k,p},x_t)+m\Big]_+,\ p\neq t \tag{7}$$

(2)当 $p=t$ 时。三元组对应的正样本为 $a_{k,p}$，并且从小批量数据中随机采样 M 帧图像作为负样本，则损失函数为：

$$L_{I\_weighted}=\Big[\sum_{x_p\in a_{k,p}}\omega_p d(x_{k,p},x_p)-\sum_{x_M\in M}\omega_M d(x_M,a_{k,t})+m\Big]_+,\ p=t \tag{8}$$

式(7)~式(8)是基于关联排序，由 $rank-1$ 判断三元组而设计的损失函数。为了挖掘轨迹中图像之间潜在的关联性，提取更鲜明的轨迹特征，根据目标图像与正负样本之间特征距离的大小来自适应加权训练模型，模型参数可由如下公式计算求得：

$$\begin{cases}\omega_p=\dfrac{e^{d(x_{k,p},x_p)}}{\sum\limits_{x\in a_{k,p}}e^{d(x_{k,p},x)}}\\ \omega_n=\dfrac{e^{-d(x_{k,p},x_t)}}{\sum\limits_{x\in a_{k,t}}e^{-d(x_{k,p},x)}}\\ \omega_M=\dfrac{e^{-d(x_M,a_{k,t})}}{\sum\limits_{x\in M}e^{-d(x,a_{k,t})}}\end{cases} \tag{9}$$

由式(9)可以看出，对于正样本，在计算 ω_p 时，困难的样本与目标样本间的特征距离大，则分配的权重会大，模型训练时会更加注重困难样本学习；而简单的样本与目标样本间的特征距离小，分配的权重也会小。对于负样本，在计算 ω_n 和 ω_M 时，困难的样本与目标样本间的特征距离小，在设计时指数变成负号，从而保证分配给困难样本的权重更大。

此外在单相机内关联学习的过程中，每个小批量样本迭代时，都要对样本集中的图像进行采样计算 $L_{I\_weighted}$，并持续更新锚样本轨迹集合，当数据集规模较大时，会造成计算资源和时间的浪费，这里采用了典型的随机梯度下降法来优化模型训练。

综上所述，这种设计的关联学习方案，在无标签数据集的前提下，可以采用一种端到端的深度学习方式。将单相机内的任意轨迹初始化为构成轨迹的帧特征的均值，以此减少计算成本，采用指数滑动平均的方法在批量迭代学习的过程中持续更新轨迹，保证轨迹特征与最近迭代的特征相关；对所有锚样本轨迹集合进行排序，确定 $rank-1$ 轨迹，并作为判断三元组的关键条件；在 $rank-1$ 轨迹的条件下，确定三元组，由此设计损失函数，并引入自适应权重挖掘样本间潜在的关联性，在批量学习中能够动态调整正负样本间的特征距离，可以加速模型的收敛速率，避免过拟合的风险，提高模型的鲁棒性。为此，这种方案能够有效学习单相机下具有判别力的轨迹

特征,从而促进跨相机下轨迹关联的效率。

1.2 挖掘跨相机三元组锚样本和损失计算

由式(2)得到单相机内的轨迹排序列表详见图3。在模型迭代过程中,采用如下方式连接 2 台摄像机 k,l 下的轨迹,作为跨相机关联学习的锚样本,即:

$$X_{k,i}^{t+1} = \frac{1}{2}(l_2(a_{k,i}^{t+1}) + l_2(a_{l,t}^{t})) \tag{10}$$

其中,$a_{k,i}$ 表示摄像机 k 中的 $rank-1$;$a_{l,t}$ 表示摄像机 l 中的 $rank-1$;t 表示样本集训练迭代的次数。

由式(10)得到跨相机锚样本集合为 $\{X_{k,i}\}_{i=1}^{N_k}$,定义跨相机的关联损失如下:

$$L_{C\_weighted} = \begin{cases} \left[D_{X_{p,p}} - \sum_{x_t \in a_{k,t}} \omega_n d(x_{k,p}, x_t) + m\right]_+ & \text{if } p \neq t \\ \left[D_{X_{p,p}} - \sum_{x_M \in M} \omega_M d(x_M, a_{k,t}) + m\right]_+ & \text{if } p = t \end{cases} \tag{11}$$

其中,$D_{X_{p,p}}$ 表示要查询的目标图像 $x_{k,p}$ 与跨相机关联的轨迹 $X_{k,p}$ 之间的特征距离,而 $X_{k,p}$ 即是由式(10)获得的与源轨迹 $a_{k,p}$ 关联的轨迹特征。ω_n 与 ω_M 即是由式(9)获得。这种三元组损失函数将会有助于该深度模型推进跨相机下最匹配的轨迹合并成含有丰富信息的跨相机锚样本,并且此种关联的轨迹特征将有效对应于要寻找的目标图像特征。

1.3 联合优化关联损失

在模型训练中,还要知道模型识别的差异,通过联合单相机关联损失 $L_{I\_weighted}$ 与跨相机关联损失 $L_{C\_weighted}$ 作为模型训练的最终损失,数学计算公式为:

$$L = L_{I\_weighted} + \lambda L_{C\_weighted} \tag{12}$$

其中,λ 是一个平衡参数。

在模型训练中,单相机内的轨迹特征学习见图2。随着模型的训练更新,要搜索的目标图像与源轨迹之间的关联程度更深,能够有效判别轨迹,从而增强跨相机下轨迹的关联程度,有效提高跨相机内的关联学习。因此,为了使模型对 2 种关联学习的程度一致,这里将 λ 设置为 1。

2 实验结果和分析

2.1 实验设置

本文采用标准视频数据集 iLIDS-VID[23]、PRID2011[24] 和 MARS[10] 来评估算法模型。文中的数据参数见表 1。

表 1 数据集参数

Tab. 1 Parameters for the datasets

数据集	IDs	Train	Test	Cameras	Track-lets	Produced
MARS	1 261	625	636	6	20 478	DPM+GMMCP
iLIDS-VID	300	150	150	2	600	hand
PRID2011	178	89	89	2	1 134	hand

在 MARS 数据集中共有20 478条行人轨迹,包括1 261个行人,每个行人至少穿过 2 台摄像机视野。在 6 台摄像机部署的监控环境下采集的行人轨迹更加贴近实际的监控场景,包含更多的未知变化。在 iLIDS-VID 数据集中共有 300 个行人,包含 600 条轨迹,在不同的摄像机下共有 2 条轨迹,每条轨迹由 23~192 张不等的连续图像构成,平均会有 73 张图像。在 PRID2011 数据集中共有 178 个行人,包含1 134条轨迹,每条轨迹由 5~675 帧图像构成。

本文中,将 MARS 数据集中的 625 个行人的轨迹用来训练,其余的 636 个行人的轨迹用来测试模型。将 iLIDS-VID 中的行人平均划分作为训练集和测试集。对于 PRID2011,采用传统的分割方案,将 178 个行人平均划分用来训练和测试,每条轨迹至少包含 27 帧图像。

本文中采用累积匹配特性 *CMC* 值来评估基于 iLIDS-VID 和 PRID2011 算法的性能,学习过程中将行人标签随机划分,重复 10 次,确保统计结果稳定。采用 *CMC* 和平均精度均值 *map* 来评估基于 MARS 算法的性能。

仿真实验是基于 Linux 系统,搭建 GPU 版的 Tensorflow[25] 框架,使用 Python 编写完成的。利用基于 ImageNet[26] 预训练的参数初始化该深度模型。为了保证采样的小批量集中都包含所有摄像机下的行人,将 *batch_size* 设置为 128。对于较大规模的数据集 MARS,设置迭代次数为 2×10^5,并采用随机梯度下降(Stochastic Gradient Descent,SGD)的方法训练模型。将初始化学习率设置为 0.01,当模型迭代剩下 5×10^4时,学习率下降为 0.001。自适应加权训练模型,为了避免被零除,在实验中,将权重衰减速率设为 e^{-6}。对于较小规模数据集 iLIDS-VID 和 PRID2011,将学习率初始为 0.045,设置迭代次数为 4×10^4,采用 RMSProp 优化器[27] 优化模型时,设置指数衰减为每 2 个 *epoches* 为 0.94。此外,则根据经验将 2 种关联损失的阈值 m 设为 0.2。在测试阶段,研究获取的轨迹特征是遵循 l_2 标准化。对跨相机下轨迹间的 l_2 距离进行计算,作为相似度测量的

标准，用于视频行人重识别中。

2.2　结果和分析

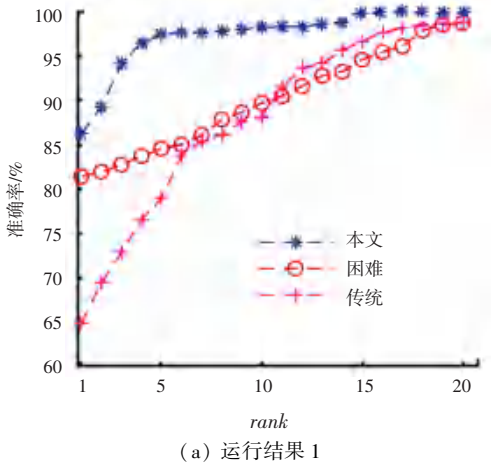
基于 ImageNet 预训练参数来初始化本文模型，采用典型的 MobileNet[28] 网络作为本文模型的骨干网络。对此过程可给出探讨论述如下。

(1)本文设计的自适应加权损失与其它损失对比。为证明本文优化的自适应加权三元组损失函数能够有效提高模型的准确度，基于标准数据集 PRID2011、iLIDS-VID 和 MARS(这里的各数据集皆为 *rank* - 1 轨迹)，与使用权重一致的传统三元组损失函数和困难样本权重的三元组损失函数做对比，说明本文采用自适应加权的方法更适用于行人重识别研究。比较结果见表 2，*CMC* 曲线如图 4 所示。

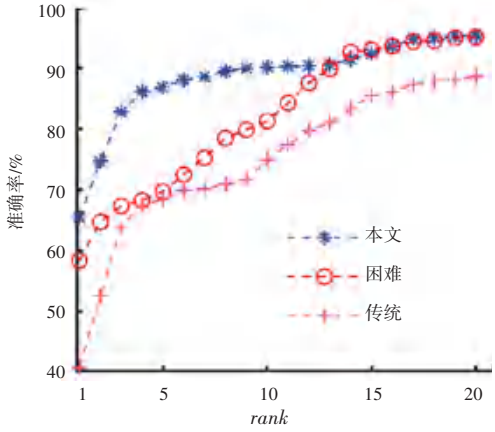
表 2　不同关联损失之间的比较

Tab. 2　Comparisons between different association loss

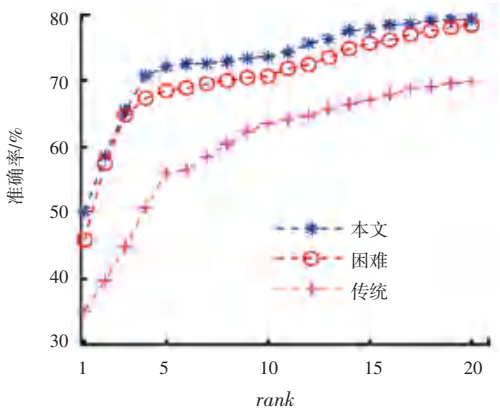
损失函数	数据集		
	MARS *rank* - 1	iLIDS-VID *rank* - 1	PRID2011 *rank* - 1
传统三元组损失	64.8	40.6	35.1
困难三元组损失	81.4	58.3	45.9
本文损失	86.2	65.4	50.2



(a) 运行结果 1



(b) 运行结果 2



(c) 运行结果 3

图 4　基于不同数据集的 3 种损失性能比较

Fig. 4　Comparison of three loss performance based on different datasets

实验证明，本文引入自适应权重，动态训练模型，提高模型的准确度更有效。由表 2 可以看出，本文模型基于 3 种标准数据集训练结果均比使用传统和困难三元组损失高。在 MARS 这种多摄像头捕捉、更贴近于现实监控场景中，本文 *rank* - 1 相较于其它 2 种损失分别高出 4.3% 和 15.1%。在数据集 iLIDS-VID 和 PRID2011 上，本文 *rank* - 1 比另外 2 种损失分别高出 7.1% 和 24.8% 以及 4.3% 和 15.1%。再结合基于不同数据集的 3 种损失性能比较的 *CMC* 曲线图如图 5 所示，图 5 中的蓝色曲线是本文模型性能。从图 5 中可以直观看出，基于本文设计的损失函数的模型性能明显优于另外 2 种损失，在不使用任何行人的先验信息条件下，本文的 *rank* - 1 基本可以达到 50% 以上。

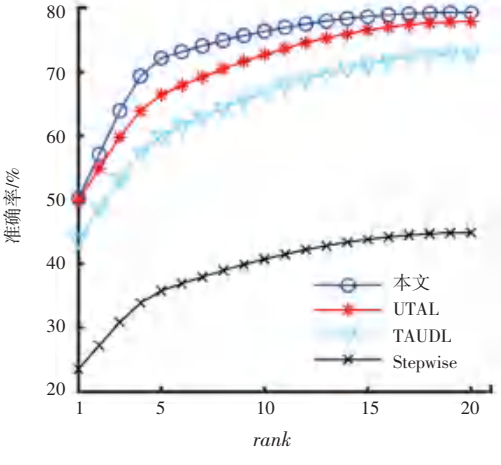


图 5　基于 MARS 的 *CMC* 曲线图

Fig. 5　*CMC* curve on MARS

(2)本文算法与其它较先进算法对比。本文中先基于较大数据集 MARS 进行实验，分别与 2020 年

较先进的算法 UTAL[29]、以及其它较先进的算法 Stepwise[18]等做比较,比较结果见表 3。

表 3 在 MARS 上的结果比较

Tab. 3 Comparison results on MARS %

算法	MARS			
	rank - 1	*rank* - 1	*rank* - 1	*map*
Stepwise	23.6	35.8	44.9	21.3
TAUDL	43.8	59.9	72.8	29.1
UTAL	49.9	66.4	77.8	35.2
本文	50.2	72.1	79.3	22.9

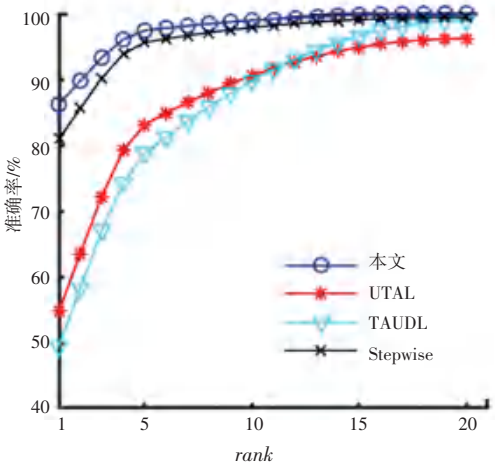
实验证明,在选用了较大的数据集、且更加接近真实的监控场景中,本文模型识别的准确率明显优于其它模型。本文算法的 *rank* - 1 为 50.2%,要比先进的 UTAL 算法 *rank* - 1 高出 0.3%。这就说明本文模型在没有任何先验行人信息的前提下,更加适用于行人重识别任务。

此外本文在标准的较小数据集 PRID2011 和 iLIDS-VID 上做了对比实验。实验结果见表 4。CMC 曲线如图 6 所示。

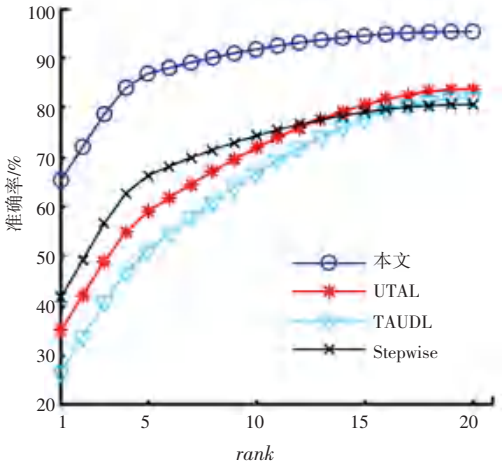
表 4 在 PRID2011 和 iLIDS-VID 上的结果比较

Tab. 4 Comparison results on PRID2011 and iLIDS-VID %

算法	PRID2011			iLIDS-VID		
	rank - 1	*rank* - 5	*rank* - 20	*rank* - 1	*rank* - 5	*rank* - 20
DVDL	40.6	69.7	85.6	25.9	48.2	68.9
UnKISS	59.2	81.7	96.1	38.2	65.7	84.1
Stepwise	80.9	95.6	99.4	41.7	66.3	80.7
TAUDL	49.4	78.7	98.9	26.7	51.3	82.0
UTAL	54.7	83.1	96.2	35.1	59.0	83.8
本文	86.2	97.4	100.0	65.4	86.9	95.4



(a) PRID2011



(b) iLIDS-VID

图 6 基于 PRID2011 和 iLIDS-VID 的 *CMC* 曲线图

Fig. 6 *CMC* curve based on PRID2011 and iLIDS-VID

实验证明,在较小的数据集上,本文算法识别准确率更高, *rank* - 1 分别为 86.2%,65.4%,相较先进的 Stepwise[16]算法分别高出了 5.3%和 23.7%。在图 6 中蓝色曲线代表本文的算法,可以直观看出比其它较先进的算法高出较多,模型性能更好,在无监督学习条件下,基于 PRID2011 训练的模型准确率达到 85%以上。基于 iLIDS-VID 数据集训练的模型性能,从图 6 中也可以看出明显高于其它算法性能, *rank* - 1 比黑色曲线高出 23.7%。

在结合不同损失函数性能对比和与当前较先进算法的比较中可以发现,本文算法较优越主要可归因为基于 *rank* - 1 挖掘的三元组较困难。具体地,当 *rank* - 1 轨迹不是源轨迹时,表明该轨迹是与目标样本距离最近的负样本、即困难样本;当 *rank* - 1 轨迹是源轨迹时,本文随机采样的 *M* 张图像作为负样本,再通过图像间的特征距离来分配权重,对困难样本着重学习。而在特征学习的过程中,基于困难三元组学习可以得到更加有效的特征。综上所述,本文模型在不使用任何先验身份信息的前提下,更加适用于行人重识别任务。

3 结束语

本文提出无监督学习三元组用于视频行人重识别研究。在基于单相机内轨迹的时空一致性学习轨迹特征过程中,利用关联排序的方法从无标签的数据集中挖掘目标图像的三元组用于计算损失,并引入自适应加权的方法来动态调整正负样本间的距离,提高模型的鲁棒性,学习单相机下具有判别力的行人特征。同时基于 *rank* - 1 合并 2 台不同摄像机下的关联轨迹,作为跨相机损失计算的三元组锚样本。最终

联合 2 种关联损失优化，提高无监督模型的准确度。

参考文献

[1] 唐佳敏，韩华，黄丽，等. 无监督行人重识别的判别性特征研究[J]. 智能计算机与应用，2021，11(08)：146-150.

[2] LI Wei，ZHU Xiatian，GONG Shaogang. Harmonious attention network for person re-identification[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City：IEEE，2018：2285-2294.

[3] HAN Hua，MA Wenjin，ZHOU Mengchu，et al. A novel semi-supervised learning approach to pedestrian re-identification[J]. IEEE Internet of Things Journal，2021，8(4)：3042-3052.

[4] ZHONGZhun，ZHENG Liang，ZHENG Zhedong，et al. Camera style adaptation for person re-identification[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City：IEEE，2018：5157-5166.

[5] 张仕远，丁学明. 融合损失优化的行人重识别方法[J]. 智能计算机与应用，2021，11(04)：65-71.

[6] HANHua，ZHOU Mengchu，SHANG Xiwu，et al. KISS+ for rapid and accurate pedestrian re-identification[J]. IEEE Transactions on Intelligent Transportation Systems，2021，22(1)：394-403.

[7] HANHua，ZHOU Mengchu，ZHANG Yujin . Can virtual samples solve small sample size problem of KISSME in pedestrian re-identification of smart transportation?[J]. IEEE Transactions on Intelligent Transportation Systems，2020，21(9)：3766-3776.

[8] ZHANG Tao，YI Zhengming，LI Xuan，et al. Improved algorithm for person re-identification based on global features[J]. Laser & Optoelectronics Progress，2020，57(24)：241503.

[9] YOU Jinjie，WU Ancong，LI Xiang，et al. Top-push video-based person re-identification[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas，NV，USA：IEEE，2016：1345-1353.

[10] ZHENG Liang，BIE Zhi，SUN Yifan，et al. Mars：A video benchmark for large-scale person re-identification[C]//European Conference on Computer Vision. Amsterdam，Netherlands：Springer International Publishing，2016，9910：865-884.

[11] LI Minxian，ZHU Xiatian，GONG Shaogang. Unsupervised person re-identification by deep learning tracklet association[M]//FERRARI V，HEBERT M，SMINCHISESCU C，et al. Computer Vision - ECCV 2018. ECCV 2018. Lecture Notes in Computer Science( ). Cham：Springer，2018，11208：772-788.

[12] ZHU Xiaoke，JING Xiaoyuan，YOU Xinge，et al. Video-based person re-identification by simultaneously learning intra-video and inter-video distance metrics[J]. IEEE Transactions on Image Processing，2018，27：5683-5695.

[13] MA Xiaolong，ZHU Xiatian，Gong Shaogang，et al. Person re-identification by unsupervised video matching[J]. Pattern Recognition，2017，65(C)：197-210.

[14] WANG Taiqing，GONG Shaogang，ZHU Xiatian，et al. Person re-identification by discriminative selection in video ranking[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2016，38(12)：2501-2514.

[15] MA Wenjin，HAN Hua，KONG Yong，et al. A new date-balanced method based on adaptive asymmetric and diversity regularization in person re-identification[J]. International Journal of Pattern Recognition and Artificial Intelligence，2020，34(9)：2056004.

[16] WANG Chunhui，HAN Hua，SHANG Xiwu，et al. A new deep learning method based on unsupervised domain adaptation and re-ranking in person re-identification[J]. International Journal of pattern Recognition and Artificial Intelligence，2020，34(13)：1-20.

[17] YE Mang，MA A J，ZHENG Liang，Jiawei Li，et al. Dynamic label graph matching for unsupervised video re-identification[C]//2017 IEEE International Conference on Computer Vision (ICCV). Venice，Italy：IEEE，2017：5152-5160.

[18] LIU Zimo，WANG Dong，LU Huchuan. Stepwise metric promotion for unsupervised video person re-identification[C]//2017 IEEE International Conference on Computer Vision (ICCV). Venice，Italy：IEEE，2017：2448-2457.

[19] COURVILLEB Y，VINCENT P. Representation learning：A review and new perspectives[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2013，35(8)：1798-1828.

[20] MCLAUGHLIN N，RINCON J M D，MILLER P. Recurrent Convolutional Network for video-based person re-identification[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas，NV USA：IEEE，2016：1325-1334.

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

[22] SCHROFF F，KALENICHENKO D，PHILBIN J. FaceNet：A unified embedding for face recognition and clustering[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston，MA，USA：IEEE，2015：815-823.

[23] WANG Taiqing，GONG Shaogang，ZHU Xiatian，et al. Person re-identification by video ranking[J]//13th European Conference on Computer Vision (ECCV). Zurich ，Switzerland：Springer，2014：688-703.

[24] HIRZER M，BELEZNAI C，ROTH P M，et al. Person re-identification by descriptive and discriminative classification[C]//Lecture Notes in Computer Science. Ystad，Sweden：Swedish Soc Automated Image Anal，2011：91-102.

[25] ABADI M，BARHAM P，CHEN Jianmin，et al. Tensorflow：A system for large-scale machine learning[J]. OSDI´16：Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation. Berkeley，CA United States：ACM，2016：265-283.

[26] DENG Jia，DONG Wei，SOCHER R，et al. ImageNet：A large-scale hierarchical image database[C]//Proceedings / CVPR，IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Miami Beach，FL：IEEE，2009：248-255.

[27] TIELMAN T，HINTON G. Lecture 6.5-rmsprop：Divide the gradient by a running average of its recent magnitude[J]. COURSERA：Neural Networks for Machine Learning，2012，4(2)：26-31.

[28] HOWARDA G，ZHU Menglong，CHEN Bo，et al. Mobilenets：Efficient convolutional neural networks for mobile vision applications[J]. ArXiv preprint ArXiv：1704.04861 ，2017.

[29] LI Minxian. ZHU Xiatian，GONG Shaogang. Unsupervised tracklet person re-identification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence 2020，42(7)：1770-1782.