ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 39,No. 5,Sep. 2024,pp. 1192－1203
DOI:10. 16337/j. 1004-9037. 2024. 05. 011

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于多核扩展卷积的无监督视频行人重识别

刘仲民[1,2]，张长凯[1,2]，胡文瑾[3]

(1. 兰州理工大学电气工程与信息工程学院,兰州 730050;2. 甘肃省工业过程先进控制重点实验室，兰州 730050;
3. 西北民族大学数学与计算机科学学院，兰州 730030)

摘　要：行人重识别旨在跨监控摄像头下检索出特定的行人目标。由于存在姿态变化、物体遮挡和背景干扰的不同成像条件等问题,导致行人特征提取不充分。本文提出一种利用多核扩展卷积的无监督视频行人重识别方法,使得提取到的行人特征能够更全面、更准确地表达个体差异和特征信息。首先,采用预训练的ResNet50作为编码器,为了进一步提升编码器的特征提取能力,引入了多核扩展卷积模块,通过增加卷积核的感受野,使得网络能够更有效地捕获到局部和全局的特征信息,从而更全面地描述行人的外貌特征;其次,通过解码器将高级语义信息还原为更为底层的特征表示,从而增强特征表示,提高系统在复杂成像条件下的性能;最后,在解码器的输出中引入多尺度特征融合模块融合相邻层中的特征,进一步减少不同特征通道层之间的语义差距,以产生更鲁棒的特征表示。在3个主流数据集上进行离线实验,结果表明该方法在准确性和鲁棒性上均取得了显著的改进。

关键词：行人重识别;多核扩展卷积;无监督学习;特征提取;注意力机制
中图分类号：TP391　　**文献标志码**：A

Unsupervised Video Person Re-identification Based on Multiple Kernel Dilated Convolution

LIU Zhongmin[1,2], ZHANG Changkai[1,2], HU Wenjin[3]

(1. School of Electrical Engineering and Information Engineering, Lanzhou University of Technology, Lanzhou 730050, China;
2. Key Laboratory of Gansu Advanced Control for Industrial Processes, Lanzhou 730050, China; 3. College of Mathematics and Computer Science, Northwest Minzu University, Lanzhou 730030, China)

Abstract: Person re-identification aims to identify specific individuals across surveillance cameras, overcoming challenges such as pose variations, occlusions, and background noise that often lead to insufficient feature extraction. This paper proposes a novel unsupervised video-based person re-identification method that utilizes multi-kernel dilated convolution to provide a more comprehensive and accurate representation of individual differences and features. Initially, we employ a pre-trained ResNet50 as an encoder. To further enhance the encoder's feature extraction capability, we introduce a multiple kernel dilated convolution module. Enlarging the receptive field of convolutional kernels allows the network to more effectively capture both local and global feature information, offering a more comprehensive depiction of a person's appearance features. Subsequently, a decoder is employed to restore high-level semantic information to a more fundamental feature representation, thereby strengthening feature

基金项目:国家自然科学基金(62061042);甘肃省工业过程先进控制重点实验室开放基金项目(2022KX10)。
收稿日期:2023-11-27;**修订日期**:2024-02-26

representation and improving system performance under complex imaging conditions. Finally, a multi-scale feature fusion module is introduced in the decoder output to merge features from adjacent layers, reducing semantic gaps between different feature channel layers and generating more robust feature representations. Offline experiments are conducted on three mainstream datasets, and results show that the proposed method achieves significant improvements in both accuracy and robustness.

Key words: person re-identification; multiple kernel dilated convolution; unsupervised learning; feature extraction; attention mechanism

引　言

行人重识别(Person re-identification, ReID)主要目的是在多个监控摄像头的视频流或图像中,对特定行人进行准确的跨摄像头追踪和识别,即通过一张将要查询的行人图像,从多个监控摄像头拍摄到的大型数据语料库中找出该行人的其他图像,这在监控摄像机网络中是一项重要的实际应用[1]。近年来,基于视频ReID的研究获得了学者广泛的关注,因为视频序列可以为特定身份的人提供丰富的时间和空间信息[2]。同时,随着卷积神经网络(Convolution neural network, CNN)的发展,有监督的视频ReID逐渐得到改进。但由于相机数量的增加,将面临昂贵的人工注释和繁重的标注操作,实际应用在很大程度上受到限制,因此无监督方法受到众多研究人员的青睐。

在无监督学习的背景下,行人重识别可以大致分为以下4类:基于迁移学习的方法[3]、基于一次性学习的方法[4]、基于聚类的方法[5]和基于轨迹关联学习的方法[6]。基于迁移学习的方法也称为基于域自适应的方法,是无监督方法的一个重要分支,旨在将源域训练模型转移到目标域[7]。Huang等[8]引入领域自适应注意力模型(Domain adaptive attention model, DAAM),将特征图分离为领域共享特征图和领域特定特征图,这两个特征图分别用于提高目标域的性能和减轻域发散带来的负面影响。基于一次性学习指的是在训练样本很少,甚至只有一个的情况下,依旧能进行预测。Ye等[4]设计一种具有正则化仿射壳和流形平滑项的锚嵌入方法。Ye等[6]引入一组未标记的训练数据进行跨摄像机标签估计,并学习更好的相似性度量来动态更新拍摄到的图像。基于聚类的方法是无监督学习的长期范例。随着深度神经网络的快速发展,联合优化聚类分析和表示学习的方法得到了广泛的应用。Lin等[5]引入一种自下而上的聚类方法来联合优化卷积神经网络和无标签样本之间的关系。Ding等[9]引入一种基于离散度的标准来评估自动生成的聚类的质量,表明聚类有效性量化起着重要的作用。Wu等[10]提出一种渐进式无监督学习(Progressive unsupervised learning, PUL)框架,消除了视觉跟踪中对带注释训练视频的需要,再通过聚类方法实现特征提取模型的微调。基于轨迹关联学习的方法是目前比较新颖的无监督视频ReID框架。Wu等[11]提出了一种通过逐步学习逐步稳定地提高卷积神经网络特征表示的判别能力来利用未标记轨迹的方法。Wu等[12]设计了多个无监督学习目标,包括基于图像和基于视频的无监督ReID在统一公式中解释了轨迹帧一致性、轨迹邻域紧凑性和轨迹簇结构。这些方法通常对全局特征进行提取来实现无监督学习,而局部特征的提取对行人重识别至关重要。

相反,在有监督行人重识别的方法中,局部特征的提取已经得到了广泛的研究。Gao等[13]提出了一种姿势引导可见部分匹配方法(Pose-guided visible part matching, PVPM),首先利用部分特征汇聚的姿势引导注意力(Pose-guided attention, PGA)方法,根据人物的姿势信息,调整对局部特征的关注程度,以更有针对性地捕获有助于人物再识别的局部信息。利用更具辨别力的局部特征,其次通过姿势引导可见性预测器(Pose-guided visibility predictor, PVP),用于预测身体的某个部位是否受到遮挡。Wang等[14]提出了一种端到端特征学习策略,即多粒度网络(Multiple granularity network, MGN),该模型旨在捕捉个体外观的多个方面,使其能够创建更全面和独特的表示,从而提取全局特征/局部特征。

为了获得多粒度的局部特征表示，通过将图像分割为多个块，并在不同的局部分支中调整局部块的数量来实现。这些方法在有监督学习的背景下显著改善了行人的识别性能。然而，在无监督学习中，局部特征的提取仍需要提升。因此，受到上述有监督行人重识别方法的启发，本文通过提取全局和局部特征的方法来解决行人特征无法充分提取的问题。采用ResNet50作为主干网络，并在此基础上稍微修改以适应行人重识别任务。通过在主干网络的基础上构造一个特征感知架构来学习局部特征和全局特征，最后通过深度关联学习计算最终关联损失。然而，大多数行人图像存在许多干扰因素，导致特征提取不充分，进而影响行人识别的精度。

为解决上述问题，强化骨干网络在行人特征提取方面的能力，本文提出了一种基于多核扩展卷积(Multiple kernel dilated convolution, MKDC)的无监督视频行人重识别模型。首先，将预训练的ResNet50作为编码器，设计一种MKDC模块帮助网络学习不同尺度和方向的特征，从而更好地捕捉输入数据的复杂结构。其次，引入解码器(Decoder block, DB)，通过与编码器的相应层进行特征融合，以结合高级和低级的特征表示，提高网络在还原输入时的性能。采用多尺度特征融合(Multiscale feature fusion, MSFF)模块融合相邻层中的特征，进一步减少不同特征通道层之间的语义差距。最后，通过内关联损失和跨相机关联损失进行无监督学习。

1 本文方法

本文对uPMNet(unsupervised Part models-based network)网络模型[15]的特征提取模块进行改进，提出了一种基于MKDC的无监督视频行人重识别方法。本文方法整体框架如图1所示，该网络主要由特征提取模块、特征感知模块和无监督模块组成。

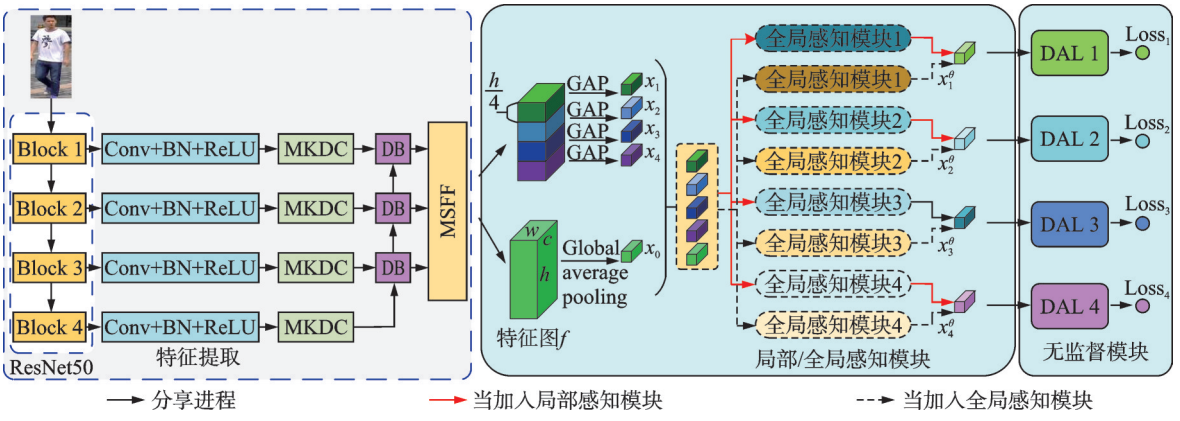

图1 本文网络整体框架

Fig.1 Overall architecture of the proposed network

将图像输入至ResNet50骨干网络，提取尺寸大小为$h\times w\times c$(h、w、c分别为高度、宽度和通道数)的初始特征f，并将f分为局部分支和全局分支。对于局部分支，将特征f按照空间分块的方式划分为4个部分，并使用全局平均池化提取局部特征$\{x_i\}_{i=1}^{k}$；对于全局分支，对特征f使用全局平均池化提取全局特征x_0，然后通过感知特征生成模块生成局部感知特征和全局感知特征。最后采用无监督方法，即深度关联学习(Deep association learning, DAL)计算最终关联损失。

在无监督模块的锚点更新阶段，引入相机内锚点I来表示每个轨迹T的特征中心，并使用部分相机内锚点I_i来表示部分轨迹T_i。使用指数移动平均值(Exponential moving average，EMA)策略递增地更新部分相机内锚点I_i，即

$$I_i^{t+1}=I_i^t-\varphi(I_i^t-x_i^\theta)\quad i=1,2,\cdots,k \tag{1}$$

式中:i表示第i部分;φ表示更新率;t表示小批量学习迭代;x_i^θ表示局部/全局感知功能。

使用全局循环排名一致性(Cyclic ranking consistency, CRC)[16]探索不同相机图像之间的关系。对于特定的I_i和$\tilde{I}_i$,它们是不同轨迹的相同部分,来自不同的相机视图。CRC意味着它们之间的欧氏距离最小,因此I_i和$\tilde{I}_i$很可能是来自不同相机的同一个人。由于I_i和$\tilde{I}_i$高度相关,引入一组部分相机交叉锚点C_i来表示它们的平均特征表示,即

$$C_i^{t+1}=\begin{cases}\dfrac{1}{2}(I_i^{t+1}+\tilde{I}_i^{t}) & \text{s.t. CRC}\\ I_i^{t+1} & \text{其他}\end{cases} \tag{2}$$

在度量学习阶段,每个小批量包含M个不同相机视角下的人物图像。每个特征x_i^θ表示一个人图像的局部/全局感知特征,此图像来自一个特定的轨迹,即其源轨迹。因此,每个特征x_i^θ有一个源部分相机内锚点I_i,计算来自同一相机视图的每个特征x_i^θ和所有部分相机内锚点I_i之间的欧氏距离。当欧氏距离最小时,意味着特征x_i^θ属于此轨迹。由于存在许多部分相机内锚点I_i,特征x_i^θ与其源部分相机内锚点I_i之间的距离表示为D_i^{I}。距离D_i^{C}可以通过特征x_i^θ及其源部分跨相机锚点C_i的相同操作获得。对于小批量中来自相同相机视图的图像,最小距离$D_i^{\min}$的平均值表示为$\bar{D}_i$,该平均值$\bar{D}_i$用于确保每个特征与其特征中心的距离相同,即源部分相机内锚点I_i。因此,上述距离是使用第i个特征及其对应的部分相机内锚点和部分跨相机锚点来计算的。

为了提高该网络学习行人特征表示的辨别能力,采用相机内关联损失和跨相机关联损失来训练网络模型,这两种损失函数被广泛地应用于各种无监督行人重识别方法。相机内关联损失L_i^{I}以强制每个帧与源轨迹的适当关联,用于判别模型学习,计算公式为

$$L_i^{\mathrm{I}}=\begin{cases}\left[D_i^{\mathrm{I}}-D_i^{\min}+m\right]_+ & D_i^{\mathrm{I}}\neq D_i^{\min}\\ \left[D_i^{\mathrm{I}}-\bar{D}_i+m\right]_+ & D_i^{\mathrm{I}}=D_i^{\min}\end{cases} \tag{3}$$

式中:$[\cdot]_+=\max(0,\cdot)$;m为给定的阈值,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;$D_i^{\min}$表示欧氏最小距离;$\bar{D}_i$表示最小距离$D_i^{\min}$的平均值。跨相机关联损失L_i^{C}能够从不相交的相机视图中学习跨相机外观变化中固有的关联轨迹,计算公式为

$$L_i^{\mathrm{C}}=\begin{cases}\left[D_i^{\mathrm{C}}-D_i^{\min}+m\right]_+ & D_i^{\mathrm{C}}\neq D_i^{\min}\\ \left[D_i^{\mathrm{C}}-\bar{D}_i+m\right]_+ & D_i^{\mathrm{C}}=D_i^{\min}\end{cases} \tag{4}$$

L_i^{I}的目标是在适当距离处将特征x_i^θ与其源部分相机内锚点I_i相关联。同时,L_i^{C}将部分相机内锚点I_i拉至接近其CRC部分相机内锚点$\tilde{I}_i$。该无监督模块的最终学习目标是联合优化两个关联损失,即

$$L_i^{\mathrm{U}}=L_i^{\mathrm{L}}+\lambda L_i^{\mathrm{C}} \tag{5}$$

式中λ为平衡这两个损失的权衡参数。

然后采用k个学习目标L_i^{U}来计算最终损失L,即

$$L=\frac{1}{k}\sum_{i=1}^{k}L_i^{\mathrm{U}} \tag{6}$$

1.1 MKDC模块

MKDC可以在特征图相同的情况下获得更大的感受野,从而获得更加密集的数据。如图2所示,MKDC模块主要由多核卷积和扩展卷积两个模块组成。多核卷积的主要目的是引入多个不同尺寸或方向的卷积核,以捕获输入数据中的多尺度和多方向的信息,而扩展卷积的作用是引入更大的感受野,捕获更广泛的上下文信息,并对特征进行更高级别的抽象。通过在多核卷积块之后添加扩展卷积,旨

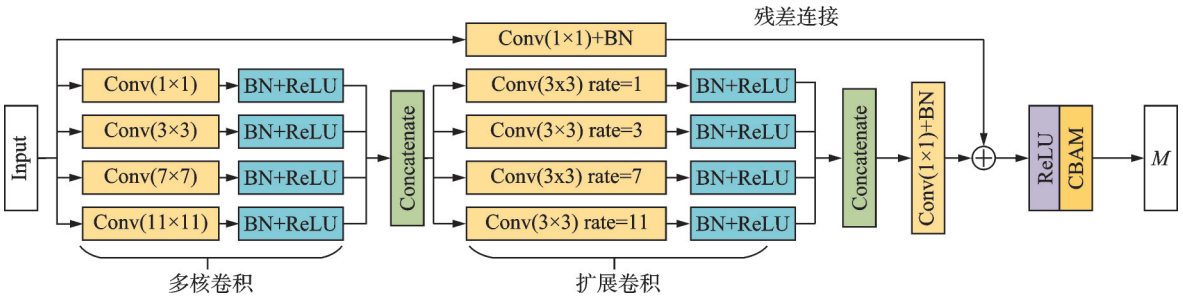

图2 MKDC模块结构图

Fig.2 MKDC structure diagram

在引入更强大的特征提取能力,提高模型的表达能力,并进一步优化对输入数据的抽象表示。

在多核卷积模块中,由4个并行卷积层开始,其卷积核大小分别为1×1、3×3、7×7和11×11,卷积核大小的逐步增加有助于捕获广泛的特性,从而使网络能够学习更鲁棒的表示。在每个卷积层之后进行批归一化和ReLU激活函数操作,然后通过1×1卷积操作(即1个全连接层)将4个不同尺寸卷积核的输出在通道维度上进行整合,将总的通道数减少到$\frac{1}{4}$,有助于减少模型的参数量和计算复杂度,同时保留重要的特征信息,最后将输出特征进行级联。

在扩展卷积模块中,通过4个卷积核都为3×3大小的并行卷积层,每个卷积层的扩张率分别为1、3、7和11,使用不同的扩展卷积有助于进一步扩大视野,允许网络捕获更多细节并细化重要特征。输出特征经过批归一化和ReLU激活函数后进行级联传播到1×1卷积后进行残差连接。最后,将生成的特征图通过CBAM(Convolutional block attention module)[17],对重要的特征赋予更高的权重,提高识别的准确性,提取深层的空间特征。经过CBAM模块处理后将特征输出,输出在图2中用M表示。

CBAM是一种轻量级端到端的注意力机制,由通道注意力模块(Channel attention module,CAM)与空间注意力模块(Spatial attention module,SAM)两类注意力机制串联组成,有效结合了两个模块的优势。模块中特征图F_i先经过通道注意力机制再经过空间注意力机制得到F_{out},如图3所示。

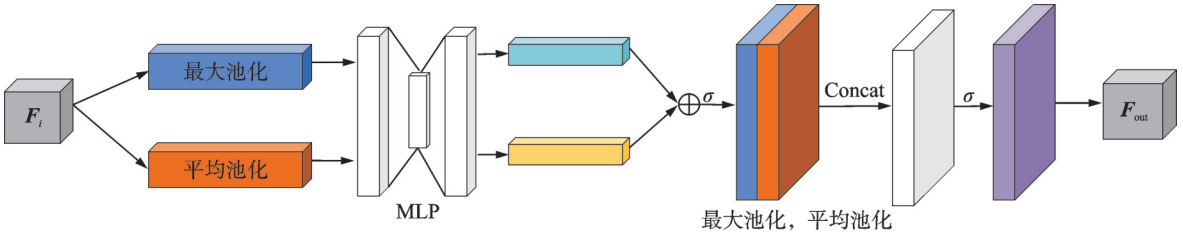

图3 CBAM结构图

Fig.3 CBAM structure diagram

在通道注意力模块中,特征图F_i分别通过最大池化及平均池化得到MaxPool(F_i)和AvgPool(F_i),然后经过两层感知机MLP(Multi-layer perceptron),将得到的输出加和生成特征图,最后将其与输入特征图F_i相乘得到,即有

$$\text{MLP}(F)=W_2(\text{ReLU}(W_1F+B_1))+B_2 \tag{7}$$

$$F_{\text{max}}^{\text{C}}=\text{MaxPool}(F_i) \tag{8}$$

$$F_{\text{avg}}^{\text{C}}=\text{AvgPool}(F_i) \tag{9}$$

$$F_{\text{CAM}_{\text{out}}}=(\sigma(\text{MLP}(F_{\text{max}}^{\text{C}})+\text{MLP}(F_{\text{avg}}^{\text{C}})))\odot F_i \tag{10}$$

式中:F为输入特征;W_1和W_2分别为两个卷积操作;B_1和B_2是偏置项;⊙代表逐点相乘;$\sigma(\cdot)$代表Sigmoid函数。

在输入空间注意力后,特征图$F_{\mathrm{CAM_{out}}}$经过最大池化和平均池化,得到两个特征图并将其按通道进行拼接,再经过一层卷积进行降维,生成注意力特征图,最后将其与输入特征图相乘,从而获得特征加权后的最终结果,即有

$$F^{\mathrm{S}}_{\max}=\mathrm{MaxPool}(F_{\mathrm{CAM_{out}}}) \tag{11}$$

$$F^{\mathrm{S}}_{\mathrm{avg}}=\mathrm{AvgPool}(F_{\mathrm{CAM_{out}}}) \tag{12}$$

$$F_{\mathrm{SAM_{out}}}=(\sigma(\mathrm{Conv}^{k\times k}(F^{\mathrm{S}}_{\max}\oplus F^{\mathrm{S}}_{\mathrm{avg}})))\odot F_{\mathrm{CAM_{out}}} \tag{13}$$

式中⊕表示级联。则总计算公式为

$$F_{\mathrm{out}}=F_{\mathrm{SAM_{out}}}(F_{\mathrm{CAM_{out}}}(F_i)) \tag{14}$$

1.2 解码器模块

为了获取更多的语义信息,增加其特征表示,采用编码-解码器。编码器将输入数据转换为潜在表示或特征,实现了对输入数据的降维和压缩,从而捕捉输入数据中的重要信息。解码器通过学习将潜在表示映射回原始数据,实现对输入数据的重建,使得输出尽可能接近输入。使用预训练的ResNet50作为编码器用于多层次特征提取,同时获取浅层的细节信息和深层的语义信息。解码器结构如图4所示。

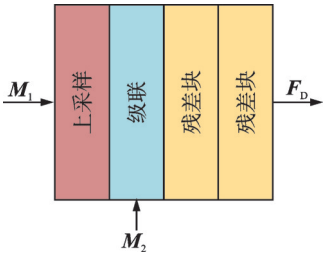

图4 解码器结构图

Fig.4 Decoder structure diagram

解码器块以双线性上采样开始,有助于将潜在表示从低维度空间还原为原始输入的高分辨率,双线性上采样通过插值技术在潜在表示的每个像素之间创建新的像素,从而实现图像的放大。之后,上采样的特征映射与另一个MKDC模块的输出进行连接,从而为解码器带来更多的语义信息,增加特征表示。为加强深度神经网络的表达能力加入了两个残差块,其中每个残差块由卷积块和连接卷积块的输入和输出的恒等映射组成。卷积块从两个3×3卷积层开始,其中每个卷积层之后都是批归一化和ReLU激活函数。

在解码器阶段,将MKDC模块的输出M_i作为解码器模块的输入后经过双线性上采样层后得到输出结果F_u,然后经过连接层与另一个MKDC块的输出M_{i+1}进行级联得到F_{C},最后将其输入至残差块生成特征图F_{D},计算公式为

$$F_u=\mathrm{upsample}(M_i) \tag{15}$$

$$F_{\mathrm{C}}=\mathrm{conv}^{k\times k}(F_u\oplus M_i) \tag{16}$$

$$F_{\mathrm{H}}=F_{\mathrm{C}}+\mathrm{Residual}(F_{\mathrm{C}}) \tag{17}$$

$$F_{\mathrm{D}}=F_{\mathrm{H}}+\mathrm{Residual}(F_{\mathrm{H}}) \tag{18}$$

式中F_{H}代表第1个残差块的输出。

1.3 MSFF模块

特征图经过多次的特征提取后,语义信息将变得十分丰富,但由于多次缩小分辨率,空间位置信息大量丢失。因此,采用多尺度特征融合的方式来得到更完备、更准确的特征信息。MSFF模块通过融合不同尺度特征图,获得更多中间层有用信息产生鲁棒的特征表示,其结构如图5所示。

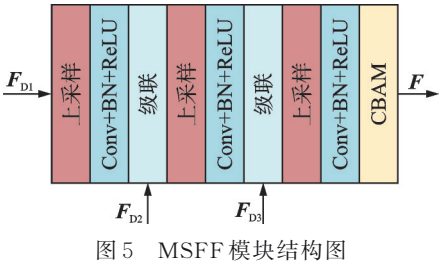

图5 MSFF模块结构图

Fig.5 MSFF structure diagram

首先,MSFF模块从第1个解码器获取输出,并将其通过双线性上采样层,将空间维度增加为原来的两倍。然后通过3×3卷积层、批归一化和ReLU激活函数操作,输出特征与第

2个解码器块的输出进行级联,通过上述相同操作,其输出特征与第3个解码器块的输出进行级联,同样进行上述操作。最后特征图通过CBAM模块,使网络更加关注具有判别力的特征,增强网络对特征的学习能力。

2 仿真实验与结果分析

2.1 数据集

实验在数据集PRID2011[18],iLIDS-VID[19]和DukeMTMC-VideoReID[20]上进行验证,如表1所示。PRID2011由校园内两个不重叠的监控摄像头拍摄,包括相机A中的385条轨迹和相机B中的749条轨迹。在这个数据集中,两个相机视图中都出现了200个人,输出了400条轨迹,平均长度为100帧。遵循文献[9]的实验设置,本文仅使用178个标识,其中每个轨迹有超过27帧。iLIDS-VID是从机场大厅两个不相交的摄像头收集,由300人的600条轨迹组成。每个人由两个来自不同相机视图的序列,其中每个轨迹的平均持续时间为73帧。DukeMTMC-VideoReID总共包含4 832个轨迹和1 812个身份,每个轨迹平均有168帧。

表1 数据集描述

Table 1 Dataset description

数据集	摄像机数	ID数	轨迹数	平均长度/帧
PRID2011	2	200	400	100
iLIDS-VID	2	300	600	73
DukeMTMC-VideoReID	8	1 812	4 832	168

2.2 实验设置

实验在NVIDIA RTX 3090Ti GPU上使用Tensorflow实现。对于PRID2011和iLIDS-VID,初始学习率为4.5×10^{-2}的优化算法RMSProp分别用于训练模型2×10^4和1.5×10^4次迭代。对于DukeMTMC-VideoReID,使用学习率初始化为1×10^{-2}的标准随机梯度下降(Stochastic gradient descent,SGD)和动量来训练模型2.5×10^4次迭代。对于所有数据集,batch_size设置为64,人物图像的大小调整为256×128,更新率φ和裕度m都设置为0.5,权衡参数λ设置为1。对于小规模数据集PRID2011和iLIDS-VID,将整套轨迹对随机分成两半,在多个试验中进行训练和测试。来自两个相机视图的同一人的轨迹分别构成了查询集和参考集。试验重复10次,以确保统计结果稳定。对于大规模数据集DukeMTMC-VideoReID,训练和测试分割方式遵循文献[11]设置。

2.3 实验结果

本文使用Rank-n和均值平均精度(mean Average precision, mAP)两种评价指标对实验结果进行性能评估。Rank-n表示识别结果中识别精度最高的前n张图片,然后计算n张图片中含有正确行人图片的百分比,从而得到行人识别的精度。mAP首先需要计算每一个物体类别的平均精度(Average precision, AP)。平均精度计算公式为

$$\mathrm{AP}=\frac{\sum P_k}{N} \tag{19}$$

式中:$\sum P_k$表示验证集中第C类目标所有精确率的和,N_c表示含有第C个类目标的图像数量。平均精度均值为每一个类别的平均精度的均值,计算公式为

$$\mathrm{mAP}=\frac{\sum_{k=0}^{C}\mathrm{AP}_k}{C} \tag{20}$$

式中:$\sum_{k=0}^{C}\mathrm{AP}_k$表示每一个类别的平均精度,$C$代表总类别数。

2.3.1 消融实验

为了验证MKDC模块中多核卷积和扩展卷积对行人识别精度的影响,分别在PRID2011、iLIDS-VID和DukeMTMC-VideoReID数据集上进行消融实验。各组实验所得到的Rank-1、Rank-5、Rank-10和mAP如表2所示。由实验结果可得,相比基准模型,多核卷积和扩展卷积都使评价指标得到相应的提升,多核卷积的提升效果相比扩展卷积更加明显,当多核卷积和扩展卷积串行连接,即使用MKDC模块时,评价指标相比多核卷积和扩展卷积两个模块达到最优。

表2 多核卷积和扩展卷积消融实验结果

Table 2 Results of ablation study on multiple kernel convolution and dilated convolution %

方法	PRID2011				iLIDS-VID				DukeMTMC-VideoReID			
	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP
多核卷积	92.5	98.1	98.5	85.9	68.7	86.4	88.5	57.6	82.7	92.4	93.7	75.8
扩展卷积	92.3	97.9	98.4	85.4	66.5	83.0	87.9	56.2	81.5	91.9	93.1	74.6
MKDC	**92.9**	**98.2**	**98.5**	**86.2**	**71.1**	**88.3**	**88.6**	**59.8**	**83.1**	**92.7**	**94.2**	**76.5**

为了验证MKDC模块中不同卷积核组合的有效性,在PRID2011、iLIDS-VID和DukeMTMC-VideoReID数据集上进行消融实验。不同大小的卷积核具有不同的感受野,即它们在输入图像中关注的空间范围不同。使用多个卷积核可以增加模型对于不同尺度特征的感知能力。1×1卷积核主要用于降维和增加非线性,能够保留通道数,减小计算量,并引入非线性变换;3×3卷积核是卷积神经网络中最常用的卷积核大小,它具有适度的感受野和参数数量,能够捕获图像中的局部结构;5×5和7×7卷积核用于捕获较大范围的上下文信息;9×9和11×11卷积核则是用于捕获更大范围的上下文信息。由于1×1和3×3卷积核的贡献不同,因此在MKDC模块加入1×1和3×3卷积核的基础上,本文针对5×5、7×7、9×9和11×11卷积核在MKDC模块中两两组合进行消融实验,各组实验所得到的Rank-1、Rank-5、Rank-10和mAP如表3所示。由实验结果可知,相比基准模型,加入不同的卷积核组合都使评价指标得到相应的提升,当同时引入7×7、11×11(MKDC模块)时评价指标相比其他组合达到最优。

表3 不同卷积核消融实验结果

Table 3 Results of ablation study on different convolution kernels %

卷积核	PRID2011				iLIDS-VID				DukeMTMC-VideoReID			
	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP
Baseline	92.0	97.7	98.4	85.2	63.1	81.9	87.8	56.4	81.3	91.7	92.9	74.6
5×5、9×9	92.1	97.9	98.4	85.5	64.4	82.7	88.0	56.6	81.5	91.8	93.1	76.5
5×5、11×11	92.6	98.2	98.5	85.8	67.2	85.6	88.5	58.9	82.6	92.5	93.8	76.3
7×7、9×9	92.4	97.8	98.5	85.6	65.9	84.2	88.2	57.1	81.7	92.2	93.5	75.3
7×7、11×11	**92.9**	**98.2**	**98.5**	**86.2**	**71.1**	**88.3**	**88.6**	**59.8**	**83.1**	**92.7**	**94.2**	**76.5**

为了验证改进后模型各个部件的有效性,在PRID2011、iLIDS-VID和DukeMTMC-VideoReID数据集上使用单一查询模式进行消融实验。在基准模型中分别加入MKDC模块、解码器和多尺度特征融合模块进行消融实验,3个模块都使行人的识别精度得到了相应的提升,尤其使用MKDC模块在PRID2011、iLIDS-VID和DukeMTMC-VideoReID三个主流数据集上的性能均有明显提升。在PRID2011数据集上,相比基准模型Rank-1提升了0.9%,Rank-5提升了0.5%,Rank-10提升了0.1%,mAP提升了1.0%;在iLIDS-VID数据集上,相比基准模型Rank-1提升了8.0%,Rank-5提升了6.4%,Rank-10提升了0.8%,mAP提升了3.4%;在DukeMTMC-VideoReID数据集上,相比基准模型Rank-1

提升了1.8%,Rank-5提升了1.0%,Rank-10提升了1.3%,mAP提升了1.9%。为了验证不同模块互相组合方式的有效性,对3个模块通过两两组合的方式进行实验,由评价指标可以看出3种互相组合的方式都有助于对行人识别精度的提升。由于MKDC模块能够增加卷积核的感受野,获取更多行人的关键特征,解码器能够进一步挖掘部分不显著的信息,更有助于提取行人特征。在PRID2011、iLIDS-VID和DukeMTMC-VideoReID三个主流数据集上的评价指标都有所提升。在PRID2011数据集上,相比基准模型Rank-1提升了1.2%,Rank-5提升了0.7%,Rank-10提升了0.3%,mAP提升了1.5%;在iLIDS-VID数据集上,相比基准模型Rank-1提升了8.5%,Rank-5提升了6.6%,Rank-10提升了2.5%,mAP提升了4.3%;在DukeMTMC-VideoReID数据集上,相比基准模型Rank-1提升了2.3%,Rank-5提升了1.3%,Rank-10提升了1.6%,mAP提升了2.3%。同时引入MKDC模块、解码器和多尺度特征融合模块对模型的性能有更加显著的提升,在PRID2011、iLIDS-VID和DukeMTMC-VideoReID数据集上的性能最佳。在PRID2011数据集上,相比基准模型Rank-1提升了1.6%,Rank-5提升了0.8%,Rank-10提升了0.4%,mAP提升了1.7%;在iLIDS-VID数据集上,相比基准模型Rank-1提升了10.6%,Rank-5提升了7.3%,Rank-10提升了2.9%,mAP提升了5.1%;在DukeMTMC-VideoReID数据集上,相比基准模型Rank-1提升了2.9%,Rank-5提升了1.7%,Rank-10提升了1.9%,mAP提升了2.5%。各组实验所得到的Rank-1、Rank-5、Rank-10和mAP如表4所示。

表4 在主流数据集上的消融实验结果

Table 4 Results of ablation study on mainstream datasets %

方法	PRID2011				iLIDS-VID				DukeMTMC-VideoReID			
	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP
Baseline	92.0	97.7	98.4	85.2	63.1	81.9	87.8	56.4	81.3	91.7	92.9	74.6
MKDC	92.9	98.2	98.5	86.2	71.1	88.3	88.6	59.8	83.1	92.7	94.2	76.5
DB	92.5	98.0	98.2	85.6	68.3	86.6	88.2	58.4	81.7	92.4	93.7	76.3
MSFF	92.4	97.8	97.9	85.3	64.8	84.2	87.9	56.9	81.5	92.1	93.1	75.3
MKDC+DB	93.2	98.4	98.7	86.7	71.6	88.5	90.3	60.7	83.6	93.0	94.5	76.9
MKDC+MSFF	93.0	98.3	98.5	86.1	71.2	88.4	89.7	60.3	83.5	92.8	94.4	76.7
DB+MSFF	92.6	98.1	98.2	85.9	68.7	86.8	88.5	60.1	81.8	92.6	94.3	76.5
MKDC+DB+MSFF	**93.6**	**98.5**	**98.8**	**86.9**	**73.7**	**89.2**	**90.7**	**61.5**	**84.1**	**93.4**	**94.8**	**77.1**

为了验证损失函数对模型性能的影响,在本文提出模型上分别对相机内关联损失和跨相机关联损失进行消融实验。各组实验所得到的Rank-1、Rank-5、Rank-10和mAP如表5所示。由实验结果可得,

表5 损失函数消融实验结果

Table 5 Results of ablation study on loss function %

方法	PRID2011				iLIDS-VID				DukeMTMC-VideoReID			
	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP
相机内关联损失	93.3	97.8	98.4	85.6	71.3	87.5	88.1	58.6	83.2	92.8	94.2	76.1
跨相机内关联损失	93.5	98.4	98.7	86.5	72.9	88.7	89.4	60.7	83.9	93.7	94.7	76.9
相机内关联损失+跨相机内关联损失	**93.6**	**98.5**	**98.8**	**86.9**	**73.7**	**89.2**	**90.7**	**61.5**	**84.1**	**93.4**	**94.8**	**77.1**

相比本文模型,相机内关联损失和跨相机关联损失都使评价指标得到相应的提升,当同时加入相机内关联损失和跨相机关联损失时评价指标达到最优。

为了更直观地展示本文方法的优越性,将检索结果可视化,如图6所示,其中红色框表示查询结果出错。从图6可以看出,第1~3行行人图像存在姿态变化和物体遮挡的情况,但仍能被正确匹配,说明本文方法针对此问题达到了显著的效果;第4行图像中检索目标行人存在背景行人的干扰,其中第7列图像出现了匹配错误的情况,说明针对背景干扰的问题,性能提升不够显著。但总的来说,本文模型能够有效抑制噪声干扰,关注有用信息,提取判别力的行人特征进行检索。

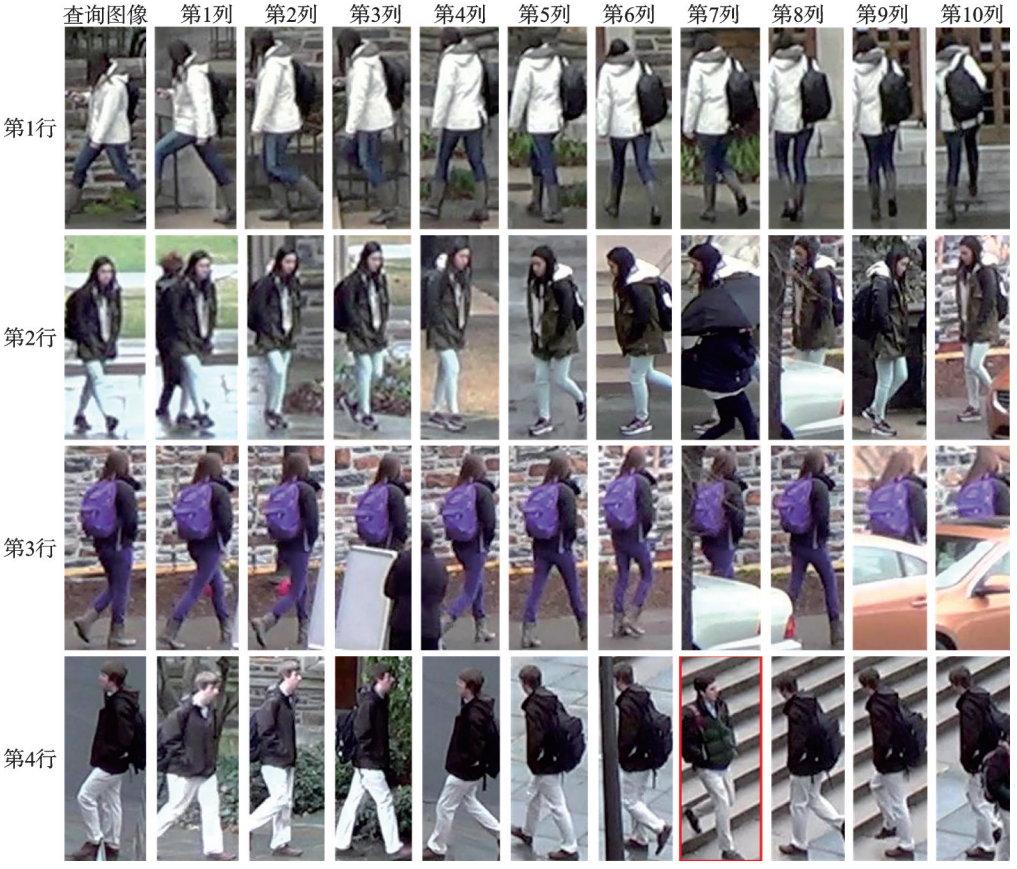

图6 检索结果可视化

Fig.6 Visualization of retrieval results

2.3.2 对比实验

为了验证本文模型的泛化性,在多个数据集上进行测试,选用PRID2011、iLIDS-VID和DukeMTMC-VideoReID主流数据集与OIM[21]、TAUDL[22]、DAL[11]、SSL[23]和uPMNet[15]、NHAC[22]等方法进行对比,实验结果如表6所示。本文方法在PRID2011数据集上Rank-1达到了93.6%,Rank-5达到了98.5%;在iLIDS-VID数据集上Rank-1达到了73.7%,Rank-5达到了89.2%;在DukeMTMC-VideoReID数据集上Rank-1达到了84.1%,Rank-5达到了93.4%,mAP达到了77.1%。与DAL、SSL和NHAC等方法相比,本文方法在3个主流数据集上的识别效果都有了明显的提升,说明本文方法能够得到更理想的实验结果,进而提取更加充分的行人特征。

表6 不同模型在主流数据集上的性能比较

Table 6 Performance comparison of different models on mainstream datasets %

方法	PRID2011		iLIDS-VID		DukeMTMC-VideoReID		
	Rank-1	Rank-5	Rank-1	Rank-5	Rank-1	Rank-5	mAP
OIM[21]	—	—	—	—	51.1	70.5	43.8
TAUDL[22]	49.4	78.7	26.7	51.3	—	—	—
DAL[11]	85.3	97.0	56.9	80.6	—	—	—
SSL[23]	—	—	—	—	76.4	88.7	69.3
NHAC[24]	—	—	—	—	82.8	92.7	76.0
uPMNet[15]	92.0	97.7	63.1	81.9	81.3	91.7	74.6
本文方法	**93.6**	**98.5**	**73.7**	**89.2**	**84.1**	**93.4**	**77.1**

注:“—”表示未得出结果。

3 结束语

本文针对无监督视频行人重识别过程中行人特征提取不充分的问题,对uPMNet网络模型特征提取模块进行改进,提出一种基于MKDC的无监督视频行人重识别方法。将预训练的ResNet50作为编码器,在编码器中引入MKDC模块增加卷积核感受野,以获取更加关键的图像信息,输出更具代表性的行人表达特征,通过解码器来获取更多的语义信息,增加其特征表示;引入多尺度特征融合模块以产生更鲁棒的特征表示来增强不同尺度的特征。在消融实验中,相比基准模型,本文模型在3个主流数据集PRID2011、iLIDS-VID和DukeMTMC-VideoReID中,Rank-1分别提升了1.6%、10.6%和2.9%,Rank-5分别提升了0.8%、7.3%和1.7%;在数据集DukeMTMC-VideoReID中mAP提升了2.5%。对比在PRID2011和iLIDS-VID数据集上Rank-n最高的DAL模型,Rank-1分别提升了8.3%和16.8%,Rank-5分别提升了1.5%和8.6%;对比在DukeMTMC-VideoReID数据集上Rank-n和mAP最高的NHAC模型,Rank-1提升了1.3%,Rank-5提升了0.7%,mAP提升了1.1%。因此,该模型在无监督视频行人重识别任务中能够实现较高的性能指标和识别精度。

参考文献:

[1] 孙明浩,王洪元,吴琳钰,等.基于特征金字塔分支和非局部关注的行人重识别[J].数据采集与处理, 2023, 38(1): 121-131.
SUN Minghao, WANG Hongyuan, WU Linyu, et al. Person re-identification based on feature Pyramid branches and no local attention [J]. Journal of Data Acquisition and Processing, 2023, 38 (1): 121-131.

[2] ZHANG Z, LAN C, ZENG W, et al. Multi-granularity reference-aided attentive feature aggregation for video-based person re-identification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE Press, 2020: 10407-10416.

[3] LIU J, ZHA Z J, CHEN D, et al. Adaptive transfer network for cross-domain person re-identification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA: IEEE Press, 2019: 7202-7211.

[4] YE M, LAN X, YUEN P C. Robust anchor embedding for unsupervised video person re-identification in the wild[C]//Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany: Springer, 2018: 170-186.

[5] LIN Y, DONG X, ZHENG L, et al. A bottom-up clustering approach to unsupervised person re-identification[C]//Proceedings of the AAAI conference on artificial intelligence. Hawaii, USA: AAAI, 2019: 8738-8745.

[6] YE M, LI J, MA A J, et al. Dynamic graph co-matching for unsupervised video-based person re-identification[J]. IEEE Transactions on Image Processing, 2019, 28(6): 2976-2990.

[7] KANG G, ZHENG L, YAN Y, et al. Deep adversarial attention alignment for unsupervised domain adaptation: The benefit of

target expectation maximization[C]//Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany: Springer, 2018: 401-416.

[8] HUANG Y, PENG P, JIN Y, et al. Domain adaptive attention model for unsupervised cross-domain person re-identification [EB/OL]. (2019-05-25). https://arxiv.org/abs/1905.10529.

[9] DING G, KHAN S, TANG Z, et al. Towards better validity: Dispersion based clustering for unsupervised person re-identification[EB/OL]. (2019-06-04). https://arxiv.org/abs/1906.01308.

[10] WU Q, WAN J, CHAN A B. Progressive unsupervised learning for visual object tracking[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE Press, 2021: 2993-3002.

[11] WU Y, LIN Y, DONG X, et al. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE Press, 2018: 5177-5186.

[12] WU G, ZHU X, GONG S. Tracklet self-supervised learning for unsupervised person re-identification[C]//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA: AAAI, 2020: 12362-12369.

[13] GAO S, WANG J, LU H, et al. Pose-guided visible part matching for occluded person reid[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE Press, 2020: 11744-11752.

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

[15] ZANG X, LI G, GAO W, et al. Exploiting robust unsupervised video person re-identification[J]. IET Image Processing, 2022, 16(3): 729-741.

[16] ZHANG Y, XIANG T, HOSPEDALES T M, et al. Deep mutual learning[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE Press, 2018: 4320-4328.

[17] WOO S, PARK J, LEE J Y, et al. CBAM: Convolutional block attention module[C]//Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany: Springer, 2018: 3-19.

[18] WANG T, GONG S G, ZHU X, et al. Person re-identification by video ranking[C]//Proceedings of European Conference on Computer Vision. Zurich, Switzerland: [s.n.], 2014: 688-703.

[19] HIRZER M, BELEZNAI C, ROTH P M, et al. Person re-identification by descriptive and discriminative classification[C]//Proceedings of Image Analysis:17th Scandinavian Conference. Ystad, Sweden: Springer, 2011: 91-102.

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

[21] XIAO T, LI S, WANG B, et al. Joint detection and identification feature learning for person search[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hawaii, USA: IEEE Press, 2017: 3415-3424.

[22] LI M, ZHU X, GONG S. Unsupervised person re-identification by deep learning tracklet association[C]//Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany: Springer, 2018: 737-753.

[23] LIN Y, XIE L, WU Y, et al. Unsupervised person re-identification via softened similarity learning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE Press, 2020: 3390-3399.

[24] XIE P, XU X, WANG Z, et al. Unsupervised video person re-identification via noise and hard frame aware clustering[C]//Proceedings of 2021 IEEE International Conference on Multimedia and Expo (ICME). [S.l.]: IEEE Press, 2021: 1-6.

作者简介:

刘仲民(1978-),**通信作者**,男,副教授,硕士生导师,研究方向:模式识别、图像修复和图像描述,E-mail:liuzhmx@163.com。

张长凯(1995-),男,硕士研究生,研究方向:行人重识别。

胡文瑾(1980-),女,教授,博士生导师,研究方向:图像修复和图像质量评价。

(编辑:刘彦东)