基于长短期时间关系网络的视频行人重识别

何智敏[1,2],钱江波[1,2*],严迪群[1,2],叶绪伦[1,2],王 翀[1,2]
(1. 宁波大学信息科学与工程学院,浙江宁波 315211;2. 浙江移动网络应用技术重点实验室,浙江宁波 315211)

摘 要: 行人重识别是计算机视觉领域中的一个重要研究方向,其目的是在不同的监控摄像头中识别并跟踪同一行人. 由于视频帧间存在多种时间关系,从这些关系中可以获取到对象的运动模式以及细粒度特征,因此视频重识别相比图像重识别拥有更丰富的时空线索,也更接近实际应用. 问题的关键是如何挖掘这些时空线索作为视频重识别的特征. 本文针对视频行人重识别问题,提出了一种基于Transformer的长短期时间关系网络(Long and Short Time Transformer,LSTT). 该网络包含长短期时间关系模块,提取重要时序信息并强化特征表示. 长期时间关系模块利用记忆线索存储每帧信息,并在每一帧建立全局联系;短期时间关系模块则考虑相邻帧之间交互,学习细粒度目标信息,提高特征表示能力. 此外,为了提高模型对不同目标特征的适配性,本文还设计了一个包含不同规格卷积核的多尺度模块. 该模块具有多种卷积感受野,能够更全面覆盖目标区域,从而进一步提高模型的泛化性能. 在MARS、MARS_DL和iLIDS-VID 3个数据集上的实验结果表明,LSTT模型性能最优.

关键词: 视频行人重识别;Transformer;长期时间关系;短期时间关系;多尺度
基金项目: 国家自然科学基金(No.62271274);宁波市科技项目(No.2024Z004,No.2023Z059)
中图分类号: TP391.41 **文献标识码:** A **文章编号:** 0372-2112(2024)08-2746-12
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20230342

Video-Based Person Re-Identification Using Long-Short Term Temporal Relationship Network

HE Zhi-min[1,2], QIAN Jiang-bo[1,2*], YAN Di-qun[1,2], YE Xu-lun[1,2], WANG Chong[1,2]
(1. *Faculty of Electrical Engineering and Computer Science*, *Ningbo University*, *Ningbo*, *Zhejiang* 315211, *China*;
2. *Zhejiang Key Laboratory of Mobile Network Application Technology*, *Ningbo*, *Zhejiang* 315211, *China*)

Abstract: Person re-identification is an important research direction in the field of computer vision, aiming to identify and track the same person across different surveillance cameras. Compared with image-based re-identification methods, the video-based re-identification method has richer temporal and spatial information, making it more efficient in real-world applications. Due to the existence of various temporal relationships between video frames, valuable information such as motion patterns and fine-grained features can be obtained. Therefore, how to effectively extract these temporal and spatial clues has become a key issue in video-based re-identification. In this paper, a long and short time Transformer (LSTT) network based on a temporal relationship is proposed to address the video-based person re-identification problem. The module includes long and short term relationship modules to extract important temporal information and enhance feature representation. The long-term relationship module stores information for each frame using a memory cue and establishes global connections for each video frame. The short-term relationship module considers interaction between adjacent frames to learn fine-grained target information and improve feature representation. Additionally, to improve the model's adaptability to different target features, a multi-scale module with convolution kernels of different sizes is designed. The module has multiple convolution receptive fields and can more comprehensively cover the target area, further improving the model's generalization performance. Experimental results on three datasets, namely MARS, MARS_DL, and iLIDS-VID, demonstrate that the LSTT model achieves state-of-the-art performance.

Key words: video-based person re-identification; Transformer; the long-term temporal relationship; the short-term temporal relationship; multi-scale module

收稿日期:2023-04-14;修回日期:2024-06-06;责任编辑:王天慧
*通讯作者:钱江波

Foundation Item(s): National Natural Science Foundation of China (No.62271274); Ningbo Science and Technology Project Grants (No.2024Z004, No.2023Z059)

1 引言

行人重识别(person Re-IDentification,ReID)是一类重要的计算机视觉任务,其目的是从大规模的图像或视频中检索给定查询图像或视频中相匹配的行人. 近年来,随着智能监控等领域的广泛应用,行人ReID引起了越来越多的关注. 根据查询的类型,行人ReID可以分为基于图像[1~7]和基于视频[8~13]两类方式. 与基于图像的行人识别相比,基于视频的行人识别具有更加丰富的时间和空间信息,这些信息可以用于减少视觉模糊性,从而提高行人ReID的鲁棒性和准确性. 本文专注于基于视频的行人ReID,因为它更贴近实际应用场景,并且具有更高的实用价值.

虽然视频行人重新识别技术已经取得了一些进展,但是目前大多数方法并没有充分利用视频数据中丰富的时空线索,限制了它们的鲁棒性和准确性. 空间线索挖掘可以利用视频中的视觉特征,如行人的外观和姿态等,来帮助确定行人身份. 而时间线索挖掘则可以利用视频中帧与帧之间的时间关系,提取行人在不同时间点的运动轨迹,从而更准确地匹配行人身份. 因此充分挖掘时空信息,可以提高视频行人ReID的鲁棒性和准确性.

针对空间线索的提取,目前的方法主要采用注意力机制[10,14,15]、基于局部特征[16~18]和多尺度[19,20]等技术进行特征增强. 虽然这些方法取得了不错的效果,但它们的本源仍存在一定局限性. 这些方法通常先使用方形卷积核提取特征,然后再加强这些特征. 然而,所提取的特征信息准确性取决于其感受野范围,而常规的方形卷积并不能很好地适应非方形的特征区域,如图1的白色背包无法被常规卷积完全覆盖. 小的卷积只能覆盖部分有效区域,而大的卷积则会带入干扰信息. 因此,对于不同规则的显著特征需要应用不同规格的卷积核来提取.

针对视频重识别任务,时间关系是视频区分图片的主要特征之一,因此对时间关系的处理具有重要意义. 现有的重识别研究大多只考虑了一种时间关系,如短期时间关系[21~23]或长期时间关系[11,24,25]. 短期时间关系关注的是详细的行人特征,例如姿态、表情、行走的步态等,这些特征对于行人重识别任务的关键性能具有重要作用. 而长期时间关系则更关注整体的运动模式,例如行人的走路速度、方向和轨迹等,这种关系模式为模型的特征学习提供了指导作用,对于缓解遮挡等问题也有着重要的作用. 这两种时间关系关注不同的信息,但都对提高网络的特征表达能力具有重要作用. 因此,为了充分利用时间线索,提高网络的性能,需要综合考虑长短期时间关系,并充分挖掘视频的时间线索.

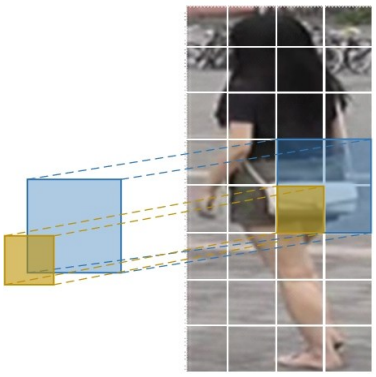
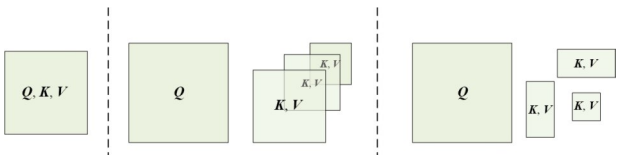


图1　常规卷积感受野示例

为了解决上述问题,本文提出一种基于Transformer的长短期时间关系模型,以充分利用视频数据的时空信息. 提出了两种时间关系模型来处理短期和长期时间关系. 对于短期时间关系,利用相邻帧之间的交互来捕捉时间关系;对于长期时间关系,采用记忆分支存储视频的运动信息,并将其与每一帧进行交互,从而在学习整体的运动信息的同时,也在每一视频帧上建立了全局联系. 特别地,使用注意力机制来实现视频帧之间的交互. 在空间特征处理方面,提出了一个多尺度模块,其中包含多种不规则(非方形)卷积核(1*1、1*2、2*1),以提高模型的特征提取能力,特别是对于一些非方形特征的目标. 与其他多尺度方法不同[26],如图2(*b*),它们的方法是利用不同尺度的方形卷积核进行特征提取,而我们的是包含多种非规则卷积核,能够更有效地满足各种目标特征. 实验结果表明,本文方法在MARS、iLIDS-VID和MARS_DL数据集上的表现优异,其性能显著超出现有方法. 这证明了本文提出的时间关系处理和空间特征处理方法的有效性,能够更好地利用视频数据中的时空信息.

综上所述,本文的主要贡献包括以下3点:

(1)提出了一种基于Transformer的长短期时间关系网络(Long and Short Time Transformer,LSTT),以充分



(*a*) ViT　(*b*) Other methods　(*c*) Ours

图2　多尺度模块对比示例

挖掘视频中的时空信息，提高行人重识别的鲁棒性和准确性. 通过大量的实验和消融实验，验证了该模型和各个模块的有效性. 在3个行人钟识别数据集上，LSTT模块均达到了最显著性能.

(2)在LSTT模块中，提出了两种时间关系模块来处理时间关系，分别为短期关系模块和长期关系模块. 这两个关系模块能够充分利用视频数据中的时序信息，从而更好地挖掘视频中的时间关系，进一步提高模型的性能.

(3)在空间特征处理方面，本研究提出了一种包含非方形卷积核的多尺度模块，提高了模型对多尺度特征的适配性和鲁棒性.

2 相关工作

本文将重点介绍基于视频的重识别方法的发展情况，主要从2D卷积、3D卷积和Transformer 3个方面进行探讨. 这些方法以视频片段作为检测单元，相比图像重识别，具有更丰富的多样化特征信息. 视频重识别具有两个优势，一是视频片段能提供更多的运动信息，二是视频片段中的时空信息能够缓解遮挡问题.

2.1 基于2D卷积网络的视频重识别

一些学者认为视频重识别方法是在图像重识别的基础上加入时间关系，而2D卷积网络是图像重识别中常用的网络模型. 在这些方法中，首先利用2D卷积网络学习视频帧内的空间信息，然后再考虑视频帧间的时序关系. 例如，Eom等人[27]提出了一种时空记忆网络，利用空间和时间记忆存储当前场景的背景信息和时序信息，对行人特征进行细化作用. Wang等人[28]则提出了一种金字塔型特征融合范式，设计了一种时空聚合模块，能够同时挖掘特征图内的空间信息和相邻特征图的时序信息，以逐级捕捉高维度特征.

2.2 基于3D卷积网络的视频重识别

还有一些学者认为使用3D卷积的方式能够更好地处理视频帧之间的关系问题，因为它可以同时挖掘视频帧的空间信息和时间信息. Hou等人[9]提出了一种时序互补学习网络，该网络提取连续视频帧的互补特征以进行视频中的行人重识别. 模型中还设有显著性擦除模块，其功能是擦除先前帧激活的部分特征来驱使模型挖掘新的互补特征，以这样的方式来获取到行人最终的整体特征表示. 此外，该作者还提出了另一种互补学习网络(Bilateral Complementary network，BiCnet)[29]. 与通过擦除来驱动特征学习的网络不同，该网络包含了两个分支. Detail Branch对原分辨率进行空间特征提取，Context Branch对下采样图像进行特征提取. 在时序关系上，模型还包含时空核选择模块(Temporal Kernel Selection，TKS)来动态学习短期和长期的时序关联. 虽然3D卷积能够同时处理视频帧的空间信息和时间信息，但是由于其需要处理的数据量更大，因此会带来更高的计算复杂度.

2.3 基于Transformer的视频重识别

基于Transformer的方法[30,31]被引入到计算机视觉领域中并取得鼓舞人心的成功. 这是因为其具有全局感知能力，更容易捕获实际有效的特征关系. 更重要的是它没有卷积网络中的下采样算子，能够保留更多的细粒度特征信息.

Zang等人[32]提出了一个基于Transformer的多方向、多尺度的金字塔模型. 为了获取细粒度的特征信息，还对patch应用不同的划分方式来生成不同方向的人体部位，意图通过Transformer中的多头自注意力机制让模型充分学习视频内的空间信息，以提高模型识别准确性. Wu等人[33]提出了一个Contextual对齐的Transformer模型，为了保留由于姿态、遮挡等因素引起的特征不对齐的空间语义信息，设计了时空偏移注意力模块以及残差位置嵌入. 作者尝试在视频帧之间进行语义对齐来寻找帧之间的时间关系，结合Transformer良好的特征提取能力，使视频重识别的准确率再创新高.

上述重识别方法都取得了较好的性能，但是在感受野适配性、时序信息挖掘等方面还有提升空间. 对此，本文提出一种基于Transformer的新网络，该网络具备全局感受野以及挖掘多种时间线索能力.

3 长短期时间关系网络

3.1 网络结构

LSTT的总体网络框架如图3所示，其能够很好地捕获空间和时间上的特征信息. 特别是对于时间维度上的信息，采用了两种时间关系来充分挖掘显著特征，而不是将图像帧视为独立单元，忽略图像帧之间的联系而直接聚合. 与其他视频重识别研究不同的是，他们的研究一般只关注其中一种时间关系，忽视了两者之间的相互补充. LSTT由5个部分组成，即转换特征维度的卷积投影层、用于提取帧内特征的Transformer网络、用于挖掘帧间关系的时间线索层、用于挖掘空间关系的多尺度模块和聚合视频帧信息的聚合层. 首先，通过卷积投影层，将每个三维的视频帧特征从$\boldsymbol{f}\in\mathbb{R}^{H\times W\times C}$重构为二维的特征嵌入$\boldsymbol{f}\in\mathbb{R}^{P\times C'}$，其中$H$、$W$、$C$、$P$、$C'$分别代表高度、宽度、通道数、patch数，以及潜在通道数. 然后，将嵌入特征传入编码器来学习帧内的显著特征. 再将编码器出来的特征送入时间线索层，利用视频帧之间的关系来对嵌入特征进行强化. 此外，在时间线索层

设计了长期时间和短期时间两种时间关系模块,能够更充分地挖掘时间维度上的显著特征. 时间线索层出来后的特征会先经过多尺度模块,增强后的特征被送入特征聚合层以生成最终的视频特征.

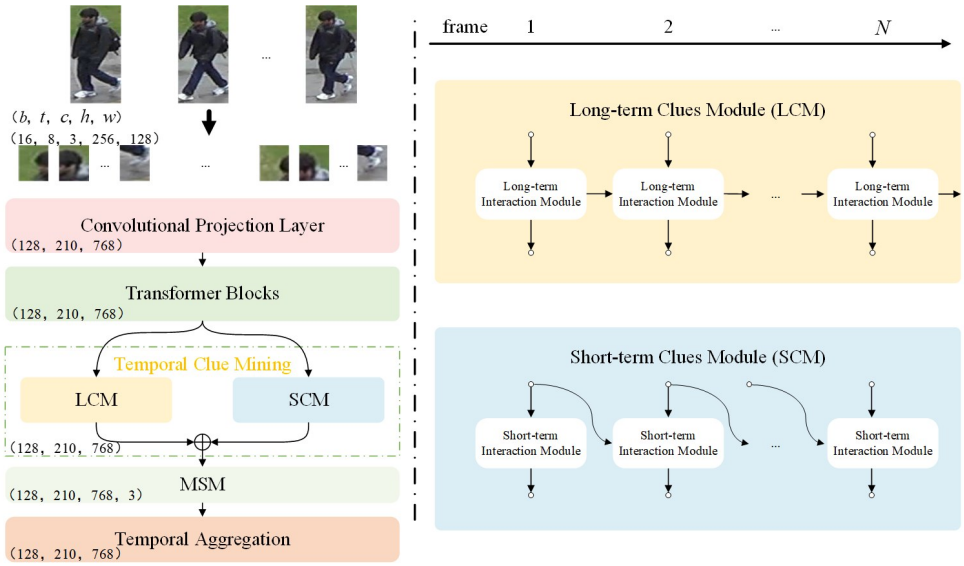


图3　LSTT网络结构

3.2　短期关系模块

将相邻视频帧之间的关系称为短期时间关系. 这种关系具有相邻帧间特征重合区域较大的特点,因此在进行短期线索挖掘时,能够更好地加强学习效果. 同时,这种短期关系的处理对于提取行人的细粒度运动信息也非常重要.

在这部分中,本文提出了短期交互模块(Short Time Interaction Module,STIM)来挖掘视频中的短期时间关系. STIM模块的结构图如图4所示,主要包含了交叉注意力模块和FFN模块. 交叉注意力模块是该模块的核心部分,它将两个相邻视频帧作为输入,并通过计算它们之间的注意力来融合两者的特征. 具体来说,STIM模块将前帧的特征向量作为$\boldsymbol{K}$、$\boldsymbol{V}$,后帧的特征向量作为$\boldsymbol{Q}$,然后进行注意力点乘运算. 同时,还构建了一个残差结构,将融合后的特征与输入时特征向量进行Add操作. 此外,FFN模块是一个全连接的前馈网络,用来提高模型的拟合能力.

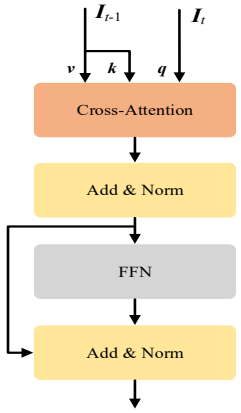


图4　短期交互模块

用$\boldsymbol{I}_t$、$\boldsymbol{I}_{t-1}$来表示第t、$(t-1)$帧的输入图像,STIM_t表示输出帧. STIM模块的数据流可由以下式表示:

$$\boldsymbol{I}_t' = \mathrm{Cross\_Attention}(\boldsymbol{I}_t, \boldsymbol{I}_{t-1}) + \boldsymbol{I}_t \tag{1}$$

$$\mathrm{Cross\_Attention}(\boldsymbol{I}_t, \boldsymbol{I}_{t-1}) = \mathrm{Attention}(\boldsymbol{I}_t, \boldsymbol{I}_{t-1}, \boldsymbol{I}_{t-1}) \tag{2}$$

$$\mathrm{STIM}_t = \mathrm{FFN}(\mathrm{LN}(\boldsymbol{I}_t')) + \boldsymbol{I}_t' \tag{3}$$

其中,FFN是由两层全连接层以及GELU激活函数构成. LN表示层归一化.

自注意力层的计算表示见式(4)~(6):

$$\mathrm{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{4}$$

$$\boldsymbol{Q} = \boldsymbol{q}^t, \boldsymbol{K} = \boldsymbol{k}^{t-1}, \boldsymbol{V} = \boldsymbol{v}^{t-1} \tag{5}$$

$$\begin{aligned} \boldsymbol{q}^t &= \boldsymbol{I}_t * \boldsymbol{W}_q \\ \boldsymbol{k}^{t-1} &= \boldsymbol{I}_{t-1} * \boldsymbol{W}_k \\ \boldsymbol{v}^{t-1} &= \boldsymbol{I}_{t-1} * \boldsymbol{W}_v \end{aligned} \tag{6}$$

其中,$\boldsymbol{W}_q$、$\boldsymbol{W}_k$以及$\boldsymbol{W}_v$是线性函数.

3.3　长期关系模块

研究表明[29],视频中的长时间关系对于探索行人运动模式非常有帮助,尤其是在遮挡问题的抑制方面起到了关键作用. 因此,本文提出了一个长期时间线索挖掘分支,由横向的记忆传输流和纵向的特征更新流构成. 记忆传输流的作用是将每个视频帧的信息进行汇聚,从而学习整段视频的运动信息和有效行人特征. 特征更新流则利用记忆传输流的信息对所有视频帧进

行交互,以强化视频帧的特征表示.

我们认为基于近距离帧所捕获到的时序关系泛化性更强,对此,记忆传输流迭代进行帧间交互捕获长期时间关系,流程见式(7):

$$\boldsymbol{I}_t^M=\mathrm{LTIM}(\boldsymbol{I}_{t-1}^M,\boldsymbol{I}_t^C),\ t=(2,3,\cdots,T) \tag{7}$$

其中,$\boldsymbol{I}_t^M$表示t时刻的记忆帧,$\boldsymbol{I}_t^C$表示t时刻的当前帧.注意$t=1$时,$\boldsymbol{I}_1^M=\boldsymbol{I}_1^C$.

长期时间交互模块(Long Term Interaction Module,LTIM)是特征更新流和记忆传输流的交互部分,其结构如图5所示.

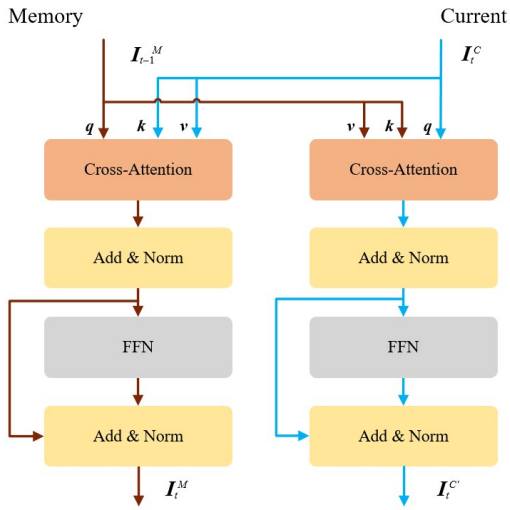


图5 长期时间交互模块

LTIM模块以记忆帧$\boldsymbol{I}_{t-1}^M$和当前帧$\boldsymbol{I}_t^C$作为输入,输出更新后的$\boldsymbol{I}_t^M$和$\boldsymbol{I}_t^{C'}$.记忆传输流和特征更新流两个分支在LTIM模块中相互作为对方注意力机制中的$\boldsymbol{K}$和$\boldsymbol{V}$.在记忆传输流中,模块以记忆帧$\boldsymbol{I}_{t-1}^M$为query,计算与当前帧$\boldsymbol{I}_t^C$的相似度,并从当前帧检索目标特征.记忆传输流计算方式见式(8)与式(9):

$$\boldsymbol{M}'=\mathrm{Cross\_Attention}(\boldsymbol{I}_{t-1}^M,\boldsymbol{I}_t^C)+\boldsymbol{I}_{t-1}^M \tag{8}$$

$$\boldsymbol{I}_t^M=\mathrm{FFN}(\mathrm{LN}(\boldsymbol{M}'))+\boldsymbol{M}' \tag{9}$$

特征更新流是LTIM模块的另一个分支,它以当前帧$\boldsymbol{I}_t^C$作为query,通过计算与key$\boldsymbol{I}_{t-1}^M$的相似度矩阵,结合value$\boldsymbol{I}_{t-1}^M$来传递特征信息.特征更新流的计算方式在式(10)与式(11)中给出:

$$\boldsymbol{C}'=\mathrm{Cross\_Attention}(\boldsymbol{I}_t^C,\boldsymbol{I}_{t-1}^M)+\boldsymbol{I}_t^C \tag{10}$$

$$\boldsymbol{I}_t^{C'}=\mathrm{FFN}\left[\mathrm{LN}(\boldsymbol{C}')\right]+\boldsymbol{C}' \tag{11}$$

3.4 多尺度模块

在网络中,将中间的特征映射为$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$向量,以便进行多头自注意力计算.在这个过程中,采用了多个独立的注意力头并行地计算自注意力.多尺度模块过程可归纳为算法1.

算法1 多尺度模块

输入: 视频帧数据$\boldsymbol{f}\in\mathbb{R}^{h\times w\times c}$;

输出: 多尺度特征$\boldsymbol{f}_{\mathrm{out}}$

1. $\boldsymbol{q}\leftarrow\mathrm{Linear}(\boldsymbol{f})$;
2. for $i\leftarrow\{1,2,3\}$ do
3. $\quad\boldsymbol{k}^i\leftarrow\mathrm{Conv}^i(\boldsymbol{f})$;
4. $\quad\boldsymbol{v}^i\leftarrow\mathrm{Conv}^i(\boldsymbol{f})$;
5. $\quad\mathrm{attn}\leftarrow\mathrm{Matmul}(\boldsymbol{k}^i,\boldsymbol{v}^i)$;
6. $\quad\mathrm{attn}'\leftarrow\mathrm{dropout}(\mathrm{attn})$;
7. $\quad\boldsymbol{f}^i\leftarrow\mathrm{Matmul}(\mathrm{attn}',\boldsymbol{q})$;
8. $\boldsymbol{f}_{\mathrm{out}}\leftarrow\mathrm{Cat}(\boldsymbol{f}^1,\boldsymbol{f}^2,\boldsymbol{f}^3)$

如图6所示,本文的多尺度模块(Multi-Scale Module,MSM)使用不同尺寸的卷积核和步长来生成不同尺度的$\boldsymbol{K}$和$\boldsymbol{V}$,以形成多尺度的特征.在同一自注意力层中,每个注意力头具有相同长度的$\boldsymbol{K}$和$\boldsymbol{V}$,而在不同的注意力头上,分别对不同尺度的$\boldsymbol{K}$和$\boldsymbol{V}$进行注意力运算,最后将多个注意力头合并以捕获不同粒度的特征信息.具体来说,对于每个注意力头$\boldsymbol{H}_i$,会将$\boldsymbol{K}$和$\boldsymbol{V}$卷积成不同大小的特征图:

$$\boldsymbol{Q}_i=\boldsymbol{X}\boldsymbol{W}_i^{\boldsymbol{Q}}, \tag{12}$$

$$\boldsymbol{K}_i,\boldsymbol{V}_i=\mathrm{MTA}(\boldsymbol{X},r_i)\boldsymbol{W}_i^{\boldsymbol{K}},\mathrm{MTA}(\boldsymbol{X},r_i)\boldsymbol{W}_i^{\boldsymbol{V}} \tag{13}$$

其中,$\mathrm{MTA}(\cdot;r_i)$表示在第i注意力头上利用第i种卷积窗口的多尺度特征聚合层.实际上,网络总共包含3种卷积窗口$r=\{(1,1),(1,2),(2,1)\}$.$\boldsymbol{W}_i^{\boldsymbol{Q}}$、$\boldsymbol{W}_i^{\boldsymbol{K}}$、$\boldsymbol{W}_i^{\boldsymbol{V}}$是第$i$注意力头上的线性映射参数矩阵.

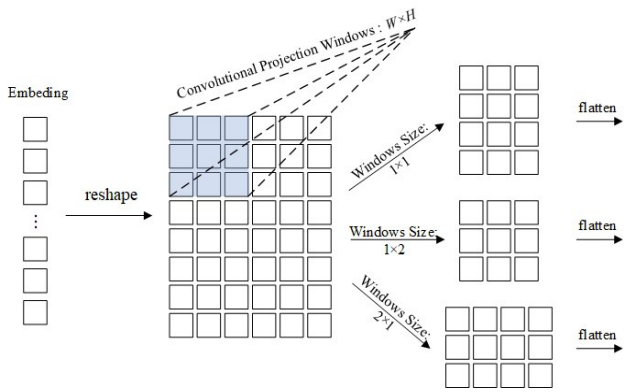


图6 多尺度模块

物体的显著特征形态多样且不规则,常规的卷积核尺寸无法准确覆盖所有特征,容易受到其他干扰特征的影响.因此,采用不同卷积核的多尺度模块进行特征提取可以更好地适应不同尺度的目标.该多尺度模块具备多种区域类型的感受野,使得模型可以更好地适应不同尺度的目标特征.同时,通过缩放特征进行卷积,可以优化模型的参数,从而提高模型的计算效率.

4　实验

4.1　数据集和评价方法

使用3个广泛使用的视频行人重识别数据集来评估本文所述的方法,包括MARS(Motion Analysis and Re-identification Set)[34],MARS_DL(re-Detect and Link on Motion Analysis and Re-identification Set)[35],iLIDS-VID(image sequence re-id dataset based on the i-LIDS MCT benchmark data)[36],这些数据集的信息被整理汇总在表1中.Liu等人[35]考虑到MARS数据集中存在很多不准确的行人边框以及错误的ID标注,利用YOLOv4重新检测了行人边界框,并使用IDE模型纠正了错误的ID标注,最终命名为MARS_DL.使用累积匹配特性(Cumulative Matching Characteristic,CMC)曲线和平均精度(mean Average Precision,mAP)来作为re-ID模型性能的评估指标.

表1　数据集信息

Dataset	ID	Tracks	Cams	Frames
MARS	1 261	20 715	6	2-920
MARS_DL	1 266	16 360	6	2-920
iLIDS-VID	300	600	2	23-192

4.2　实现细节

使用ImageNet的预训练权重来初始化Transformer编码器,选取ViT中前八层Transformer block作为主干网络.对于所有的数据集,将训练时期设置为30,并使用SGD优化器来加速模型收敛.初始学习率设为0.02,迭代器每隔1 000次学习率就缩小0.01.在训练和测试阶段,对每个视频tracklet采用限制随机采样(Restricted Random Sampling,RRS)[37]策略选取8个视频帧.所选取的视频帧的大小输入被调整为256×128,并采用随机水平翻转、填充、随机裁剪和随机擦除等数据增强方式.本文实验是在NVIDIA Tesla A100 GPU(40 GB)上进行的.

4.3　与最新方法的比较

4.3.1　MARS数据集上的实验结果与讨论

在MARS数据集上比较了近几年的最先进算法,比较结果如表2所示.LSTT在mAP指标上实现了最高的性能,在Rank-1上取得了较为优秀的性能.这些对比方法,可根据模型架构划分为2D方法和3D方法.

与3D模型的比较.与纯粹的3D卷积方法(Pseudo-3D residual net,P3D[23])相比,在mAP指标上高4.7%,在Rank-1指标上高2.2%.与时空特征对齐方法BiCent[29](Bilateral Complementary network)相比,在mAP指标上高1.9%,在Rank-1指标上高0.9%.与时空特征重构方法AP3D[23](Appearance Preserving 3D convolution)相比,在mAP指标上高2.8%,在Rank-1指标上高1.0%.对于这些性能差异,认为主要是卷积网络中局部感受野的限制导致的.同时,这些3D模型并没有很好地解决视频帧的帧间对齐问题.造成这一现象的原因有可能是,3D网络在执行时空特征学习过程中会受到过多冗余信息的干扰,导致无法对显著特征区域进行学习.

与2D模型的比较.2D模型中常见时间聚合模块有循环神经网络RNN、长短期记忆网络LSTM,以及图卷积网络GCN.最近,由于Transformer突出的特征提取能力,使得利用Transformer来捕获时间关系成为一种可能.从表中可知,相比(Spatial and Temporal Memory Networks,STMN[27]),在mAP指标上高出3.4%,在Rank-1指标上高出0.6%.相比PSTA[28](Pyramid Spatial-Temporal Aggregation),在mAP指标上高出2.1%.相比Transformer架构的DenseIL[38](Dense Interaction Learning)、PiT[32](multi-direction and multi-scale Pyramid in Transformer)和CAViT[33](Contextual Alignment Vision Transformer),在mAP指标上也有着0.9%、1.1%和0.7%的性能提升.其中,DI模型是使用ResNet50作为提取空间特征的骨干网络,并且使用Transformer进行时间建模;PiT和CAViT是使用Transformer作为特征骨干的网络.对于这些方法的性能差距,是因为这些方法没有充分考虑帧间关系,对时空信息的挖掘还不够全面.而本文的LSTT模型充分考虑了多种时间关系,这很好地解决了帧间对齐问题,充分利用了时空线索来提高模型特征表达能力.

与ResNet架构模型的比较.LSTRL[48](Long Short-Term Representation Learning)网络是在ResNet架构的基础上进行优化而成,它采用多种卷积池化策略来有效捕获长期时间关系,同时通过帧间点乘运算来捕捉短期时间依赖.与之相比,本文提出的方法在以下几个方面展现了显著的优势.首先,得益于Transformer特有的注意力机制,具有全局性的感受野,这使得模型能够直接在像素矩阵的任意位置间建立互连.这一特性使得模型在处理某一位置的数据时,能够即时纳入整个数据信息,而非局限于局部.而卷积网络受限于局部感受野,仅能捕捉到邻近区域内的信息,且需通过层级堆叠的方式逐步拓宽感受野,以期捕捉更广泛的依赖关系.但是,伴随着卷积池化操作,卷积网络容易出现数据丢失.其次,根据视频数据由连续帧构成,帧间蕴含复杂的时序关联这一特点,本模型利用交叉注意力机制更为便捷地构建起了全局的时序联系.这意味着模型能够在一个统一的视角下,有效捕获视频序列全局范围内的长期与短期依赖关系.最后,Transformer模型学习到的是全局依赖关系,这一特点使其在应对不同尺度、视角变换或遮挡等复杂情况时,展现出较卷积网

络更强的泛化能力. 这种全局性的依赖关系捕捉,为视频重识别任务提供了更为稳健和准确的处理方式.

4.3.2 iLIDS-VID数据集上的实验结果与讨论

表2为iLIDS-VID数据集上与最先进方法的比较结果. iLIDS-VID是一个小型行人视频重识别数据集,它包含300个行人ID,每个ID具有2个不同摄像头下的视频序列,共计600个视频片段. 由于该数据集视频片段少和重识别任务数据划分的原因,研究人员通常使用Rank-n作为该数据集的评估指标. 在该数据集上,LSTT在所有指标上都取得了最先进结果. 相比最近的Transformer模型PiT,在Rank-1和Rank-5上高出1.9%和0.4%. 相比利用添加时间噪声的方式提高时间相关特征的(Adversarial Feature Augmentation, AFA[39])方法,在Rank-1和Rank-5上高出5.5%和2.5%. 我们认为可能是这些方法对于时间关系的挖掘不够充分,导致模型对于视频中一些遮挡表现出较差的稳定性.

表2 各方法在MARS、iLIDS-VID数据集上的实验结果对比

方法		MARS		iLIDS-VID	
		mAP	Rank-1	Rank-1	Rank-5
2D	Multi-Granularity Reference-aided Attentive Feature Aggregation,MG-RAFA[15](CVPR' 20)	85.9	88.8	88.6	98.0
	Pompeiu-hausdorff Distance,PhD[41](CVPR' 20)	86.2	88.9	—	—
	Adaptive Graph Representation Learning,AGRL[42](TIP' 20)	81.9	89.5	84.5	96.7
	Spatial-Temporal Graph Convolutional Network,STGCN[11](CVPR' 20)	83.7	90.0	—	—
	Multi-Granular Hypergraph,MGH[43](CVPR' 20)	85.8	90.0	85.6	97.1
	Relation-Guided Temporal Refinement,RGTR[10](AAAI' 20)	84.0	89.4	86.0	98.0
	spatial-temporal Correlation and Topology Learning,CTL[44](CVPR' 21)	86.7	91.4	—	—
	Global-guided Reciprocal Learning,GRL[45](CVPR' 21)	84.8	91.0	90.4	98.3
	Pyramid Spatial-Temporal Aggregation,PSTA[28](ICCV' 21)	85.8	91.5	91.5	98.1
	Dense Interaction Learning,DenseIL[38](ICCV' 21)	87.0	90.8	92.0	98.0
	Spatial and Temporal Memory Networks,STMN[27](ICCV' 21)	84.5	90.5	—	—
	Region Feature Completion,RFCnet[47](PAMI' 21)	86.3	90.7	—	—
	Temporal-consistent Visual Clue Attentive Network,TVCAN[18](ICMR' 22)	85.7	89.0	88.5	98.1
	Contextual Alignment Vision Transformer,CAViT[33](ECCV' 22)	87.2	90.8	93.3	98.0
	multi-direction and multi-scale Pyramid in Transformer,PiT[32](ITII' 22)	86.8	90.2	92.1	98.9
	Long Short-Term Representation Learning,LSTRL[48] (ICIG' 23)	86.8	**91.6**	92.2	98.6
3D	Pseudo-3D Residual Net,P3D[23](ICCV' 17)	83.2	88.9	—	—
	Interaction Aggregation-Update Network,IAUNet[24](TNNLS' 20)	85.0	90.2	—	—
	Multi-scale 3D convolution,M3D[19](TPMAI' 20)	79.5	88.6	86.7	98.0
	Temporal Complementary Learning Network,TCLNet[9](ECCV' 20)	85.1	89.8	86.6	—
	Appearance Preserving 3D convolution,AP3D[23](ECCV' 20)	85.1	90.1	88.7	—
	Adversarial Feature Augmentation,AFA[39](ECCV' 20)	82.9	90.2	88.5	96.8
	Spatio-Temporal Representation Factorization,STRF[46](ICCV' 21)	86.1	90.3	89.3	—
	Bilateral Complementary Network,BiCnet-TKS[29](CVPR' 21)	86.0	90.2	—	—
2D	Our work	**87.9**	91.1	**94.0**	**99.3**

4.3.3 MARS_DL数据集上的实验结果与讨论

表3为MARS_DL数据集上与最先进方法的比较结果. MARS_DL是通过对MARS数据集中错误的标注类别和不标准的检测框进行筛选后重新整理的数据. 与MARS数据集一样,是一个大型的视频行人重识别数据集. 与3D方法AP3D相比,在mAP指标上有4.5%的提升,在Rank-1指标上有5.0%的提升. 与2D方法FT-WFT(Fantastic Techniques and Where to Find Them)[49]相比,在mAP指标上高出7.2%,在Rank-1指标上高出5.3%. 这两种方法是基于ResNet50作为特征提取框架,而本文的方法是基于Transformer进行时空建模. 性能差距的原因在于,Transformer具有更强的特征提取能力. 为了让模型更具说服力,对比了同Transformer架构的CAViT方法. 结果表明,LSTT在mAP和Rank-1指标上要高出0.5%和0.7%. 本文认为,LSTT性能好的原因还在于多尺度架构,这使模型对多种尺寸的特征都具有感知性. 同时,LSTT在MARS_DL数据集上的mAP和Rank-1,相比MARS数据集要高出3.1%和5.2%. 这也证实了原始数据集中错误的标注情况对模型性能影响,甚至说原来的MARS数据是无法真实地评估模型性能,也意味着在MARS_DL数据集上评估的必要性.

表3　MARS_DL数据集上的结果

方法		MARS_DL	
		mAP	Rank-1
3D	TCLNet[9](ECCV' 20)	85.4	91.0
	AP3D[23](ECCV' 20)	86.5	91.3
2D	FT-WFT[49](AAAI' 20)	83.8	91.0
	CAViT[33](ECCV' 22)	90.5	95.6
2D	Our work	**91.0**	**96.3**

4.4　消融实验

为了证明模型各模块的有效性,在MARS数据集上对长短期时间关系模块、多尺度模块进行消融实验,结果如表4所示. 可以发现,我们的基础Transformer架构就能达到85.8%的mAP和90.0%的Rank-1指标,相比表中的对比方法已经展现出一定的竞争力. 这也体现Transformer架构具有优秀的特征提取能力. 通过挖掘时间线索,模型在mAP指标上能达到87.2%的性能,在Rank-1指标上能达到90.7%的性能,达到与最先进的CAViT方法一样的性能.

表4　分析不同的组件对性能的影响

时间线索		空间线索(多尺度卷积核)	MARS	
短期时间关系	长期时间关系	方形(1) 非方形(2)	mAP	Rank-1
×	×	×	85.8	90.0
×	×	(2)	86.4	90.3
×	×	(1)	86.1	89.5
√	×	×	86.7	90.6
×	√	×	86.3	90.3
√	√	×	87.2	90.7
√	√	(1)	87.4	90.5
√	√	(2)	**87.9**	**91.1**

同时,本文还研究了哪种时间关系对模型性能贡献最大. 实验表明短期时间关系相较于长期时间关系,能够获得更高的mAP和Rank-1. 分析短期时间关系能够让模型学习到详细的细粒度信息,这对提升模型的特征表达能力具有更大帮助. 最后,在具有时间关系和多尺度空间特征增强策略时,模型达到了最高性能,87.9%的mAP和91.1%的Rank-1. 这要归功于多尺度模块的设计,其提高了模型对各种显著目标的适应性,从而进一步提高了模型的空间建模能力. 本文使用的非方形多尺度卷积相比方形卷积核,能够达到更高的mAP、Rank-1,这主要是因为非方形卷积核对各种尺寸的物体适配性更强.

如表5所示,对模型结构设置进行消融分析. 涉及结构设置有时序关系的层数和长短帧特征的聚合方式. 为了实验的公平性以及更显著的对比,在进行时间结构消融的时候,将去掉了多尺度模块. 先是探索了add、concat两种长短帧聚合方式,容易发现采取add方式能够获得更高的性能,会有0.3%的mAP提升及0.2%的Rank-1提升. 分析add方式之所以更有效,可能是因为长短期时间关系各自关注不同的特征信息(细粒度主体信息和运动模式信息),通过add能够将两种特征进行融合达到互补的作用. 然后,对时间关系的层数进行探究,实验证明二层的时间关系比一层高了0.4%的mAP和0.7%的Rank-1. 经分析,可能是一层的短期线索只能帮助两帧之间进行交互,而两层的短期线索能够使得三帧特征进行交互,扩大了模型的交互范围. 此外,在长期线索方面,两层的长期记忆流使得模型学习到的特征信息更具有显著性,且更稳定.

表5　分析时间聚合和时间层数对性能的影响

时间聚合方式	时间关系层数	MARS	
		mAP	Rank-1
Concat	1	86.3	89.6
Add	1	86.8	90.0
Concat	2	86.9	90.5
Add	2	**87.2**	**90.7**

此外,本文还在MARS_DL数据集上探究了时间关系中聚合顺序以及自注意力层中$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$的选取对模型性能的影响,实验结果如表6所示. 对于聚合顺序,即记忆流传播方向,视频帧之间信息的传播是前往后还是后往前. 由表可知,采取从前往后的信息流传播方式能够取得更好的性能,相比从后往前的传播方式,高出了0.3%的Rank-1及0.6%的mAP. 分析原因,可能是在数据集中遮挡等干扰主要存在于视频片段的后半段,而从后往前传播信息流会携带这些干扰信息,进而影响最终的整体特征表示. 本文的自注意力模块是以当前帧作为$\boldsymbol{K}$、$\boldsymbol{V}$,以相邻帧作为$\boldsymbol{Q}$,通过注意力方式将当前帧信息传给相邻帧. 对于使用当前帧作为$\boldsymbol{Q}$,而相邻帧作为$\boldsymbol{K}$、$\boldsymbol{V}$的方向也进行了验证,能达到95.9%的Rank-1和90.2%的mAP,不如本文中选取当前帧作为$\boldsymbol{K}$、$\boldsymbol{V}$的交互方式.

表6　分析时序传播方向和$\boldsymbol{K}$、$\boldsymbol{V}$选取对性能的影响

时序传播方向 $T_0 \gg T_n$	当前帧作为$\boldsymbol{K}$、$\boldsymbol{V}$,相邻帧作为$\boldsymbol{Q}$	MARS_DL	
		mAP	Rank-1
×	√	90.4	96.0
√	×	90.2	95.9
√	√	**91.0**	**96.3**

4.5　可视化分析和推演速度分析

(1)可视化分析. 为了更加清晰地呈现模型的效能,将模型的激活图输出,如图7所示. 第一行呈现了4个连续的输入帧,而第二行则展示了相应的热力激活

图. 正如图所揭示的,这些激活图主要集中在背包周围,这一现象在后面的视频帧更为显著.

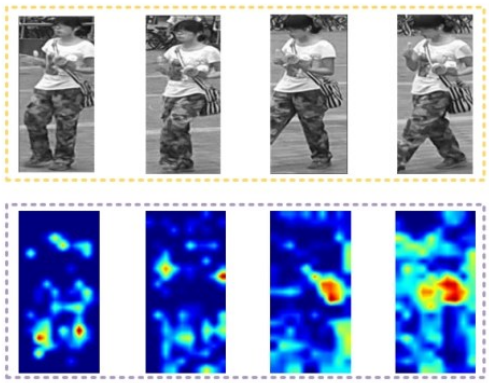


图7 热力激活图

早期帧未能充分关注背包特征,可以归结为以下3个因素. 首先,训练数据中背包样本的数量不足或存在严重不平衡,模型可能难以充分捕捉到背包的特征,更显著的区分特征往往集中在行人特征上. 其次,一些方法专门针对背包特征的独立训练,可以有效识别并利用背包的独特性来区分各个行人. 然而,本方法并未特别针对背包等属性特征进行单独设计,而是逐步自动聚焦有区分度的特征. 最后,背包的尺寸和视角对模型性能产生了显著影响. 在$\boldsymbol{I}_1$视频帧中,背包的尺寸较小,且在该视角下,背包的外观显得更为狭窄. 本研究的方法专注于从时序数据中学习关键的特征,随着视频帧序列的交互,挖掘的时间关系更为深入,从而能够学习到更为丰富的行人特征,同时也会关注到其他包括背包等有意义对象的特征.

(2)推演速度分析. 本文进一步补充了与其他方法在推演速度方面的比较实验. 选取了CAViT[33]和BiCnet-TKS[29]作为对比方法,并在MARS数据集上进行了一系列实验来评估模型的推演效率. 为了进行公正的比较,本次实验均在NVIDIA 3090显卡的服务器上完成. 实验结果如表7所示. 相较于基于Transformer架构的CAViT方法,本文方法在推理速度上取得了显著的提升. CAViT在推理过程中效率受限的原因是,其采纳的多形态Patch嵌入方法在努力捕获不同尺度的空间语义信息时,带来了附加的计算负荷,进而拖慢了处理流程的整体速度. 此外,本文方法在推理效率上也超越了基于3D卷积网络的BiCnet-TKS方法,因为后者通过3D卷积核搭建的多尺度时间关系模块对长短时间依赖进行了细致的建模导致速度较慢.

表7 MARS数据集推演速度实验分析

方法			推演总时间	每帧所需时间/s
3D		BiCnet-TKS[29](CVPR' 21)	13 min 19 s	8.83×10^{-3}
2D		CAViT[33](ECCV' 22)	46 min 1 s	3.05×10^{-2}
2D	Our work	Base	1 min 47 s	1.18×10^{-3}
		+时间线索	3 min 47 s	2.51×10^{-3}
		+时间线索 +空间线索	4 min 17 s	2.84×10^{-3}

5 结语

本文介绍了一种长短期时间关系网络,用于视频的行人重新识别任务. 该网络使用Transformer作为主干网络,具有较高的特征提取能力. 同时,采用两种时间关系进行建模,充分考虑帧间的关联,从而更好地挖掘视频序列中的时空信息,提高网络对视频序列中所存在的遮挡问题的泛化能力. 此外,还设计了一个多尺度模块,该部分与传统的多尺度不同,使用不同规格的卷积核,试图学习不同规格区域的特征,这在一定程度上提高了模型对多种特征尺度的适配性. 在多个视频行人重识别数据集上进行验证,实验结果表明本文模型展现出较好的性能.

参考文献

[1] GE Y, LI Z, ZHAO H, et al. FD-GAN: Pose-guided feature distilling GAN for robust person re-identification[J]. Advances in Neural Information Processing Systems, 2018, 31: 1-13.

[2] JIAO B, TAN X, ZHOU J, et al. Instance and pair-aware dynamic networks for re-identification[EB/OL]. (2021)[2023]. https://arxiv.org/abs/2103.05395.

[3] YE M, SHEN J, LIN G, et al. Deep learning for person re-identification: A survey and outlook[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(6): 2872-2893.

[4] ZHOU Q, FAN H, ZHENG S, et al. Graph correspondence transfer for person re-identification[C]//Proceedings of the AAAI Conference on Artificial Intelligence. New York: ACM, 2018: 3279-3288.

[5] WANG K , WANG P , DING C ,et al. Batch coherence-driven network for part-aware person re-identification[J]. IEEE Transactions on Image Processing, 2021, 30: 3405-3418.

[6] GAO Z, WEI S X, GUAN W L, et al. Identity-guided collaborative learning for cloth-changing person reidentification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(5): 2819-2837.

[7] YU Z, TIWARI P, HOU L, et al. MV-ReID: 3D multi-view transformation network for occluded person re-identi-

fication[J]. Knowledge-Based Systems, 2024, 283: 111200.

[8] FU Y, WANG X, WEI Y, et al. STA: Spatial-temporal attention for large-scale video-based person re-identification [C]//Proceedings of the AAAI Conference on Artificial Intelligence. New York: ACM, 2019: 8287-8294.

[9] HOU R B, CHANG H, MA B P, et al. Temporal complementary learning for video person re-identification[C]// Computer Vision-ECCV 2020. Cham: Springer International Publishing, 2020: 388-405.

[10] LI X, ZHOU W, ZHOU Y, et al. Relation-guided spatial attention and temporal refinement for video-based person re-identification[C]//Proceedings of the AAAI Conference on Artificial Intelligence. New York: ACM, 2020: 11434-11441.

[11] YANG J R, ZHENG W S, YANG Q Z, et al. Spatial-temporal graph convolutional network for video-based person re-identification[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 3289-3299.

[12] HU X, WEI D, WANG Z, et al. Hypergraph video pedestrian re-identification based on posture structure relationship and action constraints[J]. Pattern Recognition, 2021, 111: 107688.

[13] MEKHAZNI D, DUFAU M, DESROSIERS C, et al. Camera alignment and weighted contrastive learning for domain adaptation in video person ReID[C]//2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Piscataway: IEEE, 2023: 1624-1633.

[14] SI J L, ZHANG H G, LI C G, et al. Dual attention matching network for context-aware feature sequence based person re-identification[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 5363-5372.

[15] ZHANG Z Z, LAN C L, ZENG W J, et al. Multi-granularity reference-aided attentive feature aggregation for video-based person re-identification[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 10407-10416.

[16] SARFRAZ M S, SCHUMANN A, EBERLE A, et al. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 420-429.

[17] LIU J X, NI B B, YAN Y C, et al. Pose transferrable person re-identification[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 4099-4108.

[18] JIAO B L, GAO L Y, WANG P. Temporal-consistent visual clue attentive network for video-based person re-identification[C]//Proceedings of the 2022 International Conference on Multimedia Retrieval. New York: ACM, 2022: 72-80.

[19] LI J N, ZHANG S L, HUANG T J. Multi-scale 3D convolution network for video based person re-identification [C]//Proceedings of the AAAI Conference on Artificial Intelligence. New York: ACM, 2019: 8618-8625.

[20] MA Y, BAI T, ZHANG W Y, et al. Multi-scale relation network for person re-identification[C]//2021 IEEE Symposium on Computers and Communications (ISCC). Piscataway: IEEE, 2021: 1-7.

[21] MCLAUGHLIN N, MARTINEZ DEL RINCON J, MILLER P. Recurrent convolutional network for video-based person re-identification[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2016: 1325-1334.

[22] YAN Y, NI B, SONG Z, et al. Person re-identification via recurrent feature aggregation[C]//Computer Vision-ECCV 2016: 14th European Conference. Amsterdam: Springer International Publishing, 2016: 701-716.

[23] GU X, CHANG H, MA B, et al. Appearance-preserving 3d convolution for video-based person re-identification [C]//Computer Vision-ECCV 2020: 16th European Conference. Glasgow: Springer International Publishing, 2020: 228-243.

[24] HOU R, MA B, CHANG H, et al. IAUnet: Global context-aware feature learning for person reidentification[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 32(10): 4460-4474.

[25] WANG X L, GIRSHICK R, GUPTA A, et al. Non-local neural networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 7794-7803.

[26] REN S C, ZHOU D Q, HE S F, et al. Shunted self-attention via multi-scale token aggregation[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 10853-10862.

[27] EOM C, LEE G, LEE J, et al. Video-based person re-identification with spatial and temporal memory networks [C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2021: 12036-12045.

[28] WANG Y Q, ZHANG P P, GAO S, et al. Pyramid spatial-

temporal aggregation for video-based person re-identification[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2021: 12026-12035.

[29] HOU R B, CHANG H, MA B P, et al. BiCnet-TKS: Learning efficient spatial-temporal representation for video person re-identification[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021: 2014-2023.

[30] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. (2020)[2023]. https://arxiv.org/abs/2010.11929.

[31] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers[C]//Computer Vision-ECCV 2020: 16th European Conference. Glasgow: Springer International Publishing, 2020: 213-229.

[32] ZANG X, LI G, GAO W. Multidirection and multiscale pyramid in transformer for video-based pedestrian retrieval[J]. IEEE Transactions on Industrial Informatics, 2022, 18(12): 8776-8785.

[33] WU J L, HE L X, LIU W, et al. CAViT: contextual alignment vision transformer for video object re-identification[C]//Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2022: 549-566.

[34] ZHENG L, BIE Z, SUN Y F, et al. MARS: A video benchmark for large-scale person re-identification[C]// Computer Vision-ECCV 2016: 15th European Conference. Cham: Springer International Publishing, 2016: 868-884.

[35] LIU C T, CHEN J C, CHEN C S, et al. Video-based person re-identification without bells and whistles[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Piscataway: IEEE, 2021: 1491-1500.

[36] WANG T, GONG S, ZHU X, et al. Person re-identification by video ranking[C]//Computer Vision-ECCV 2014: 13th European Conference. Zurich: Springer International Publishing, 2014: 688-703.

[37] LI S, BAK S, CARR P, et al. Diversity regularized spatiotemporal attention for video-based person re-identification [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 369-378.

[38] HE T Y, JIN X, SHEN X, et al. Dense interaction learning for video-based person re-identification[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2021: 300.

[39] CHEN G Y, RAO Y M, LU J W, et al. Temporal coherence or temporal motion: Which is more critical for video-based person re-identification?[C]///Computer Vision-ECCV 2020: 19th European Conference. Cham: Springer International Publishing, 2020: 660-676.

[40] PATHAK P, ESHRATIFAR A E, GORMISH M. Video person re-ID: Fantastic techniques and where to find them (student abstract)[C]//Proceedings of the AAAI Conference on Artificial Intelligence. New York: ACM, 2020: 13893-13894.

[41] ZHAO J N, QI F L, REN G Y, et al. PhD learning: Learning with pompeiu-hausdorff distances for video-based vehicle re-identification[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021: 2225-2235.

[42] WU Y, BOURAHLA O E F, LI X, et al. Adaptive graph representation learning for video person re-identification[J]. IEEE Transactions on Image Processing, 2020, 29: 8821-8830.

[43] YAN Y C, QIN J, CHEN J X, et al. Learning multi-granular hypergraphs for video-based person re-identification[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 2899-2908.

[44] LIU J W, ZHA Z J, WU W, et al. Spatial-temporal correlation and topology learning for person re-identification in videos[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021: 4370-4379.

[45] LIU X H, ZHANG P P, YU C Y, et al. Watching you: Global-guided reciprocal learning for video-based person re-identification[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021: 13334-13343.

[46] AICH A, ZHENG M, KARANAM S, et al. Spatio-temporal representation factorization for video-based person re-identification[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2021: 152-162.

[47] HOU R, MA B, CHANG H, et al. Feature completion for occluded person re-identification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(9): 4894-4912.

[48] LIU X H, ZHANG P P, LU H C. Video-based person re-

identification with long short-term representation learning[C]//Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2023: 55-67.

[49] PATHAK P, ESHRATIFAR A E, GORMISH M. Video person re-id: Fantastic techniques and where to find them (student abstract)[C]//Proceedings of the AAAI Conference on Artificial Intelligence. New York: ACM, 2020: 13893-13894.

作者简介

何智敏　男,1998 年 4 月出生,浙江庆元人. 宁波大学信息科学与工程学院硕士研究生. 主要研究方向为计算机视觉.
E-mail: hezhimin7028@163.com

钱江波　男,1974 年 7 月出生,浙江宁波人. 宁波大学信息科学与工程学院教授、博士生导师. 主要研究方向为计算机视觉、数据挖掘.
E-mail: qianjiangbo@nbu.edu.cn

严迪群　男,1979 年 7 月出生,浙江宁波人. 现为宁波大学信息科学与工程学院副教授. 主要研究方向为深度学习、计算机视觉.
E-mail: yandiqun@nbu.edu.cn

叶绪伦　男,分别于 2016 年和 2019 年获得宁波大学硕士和博士学位,现为宁波大学信息科学与工程学院讲师. 研究方向为贝叶斯学习、非参数聚类和凸分析.
E-mail: yexlwh@163.com

王　翀　男,1985 年 2 月出生,浙江宁波人. 现为宁波大学信息科学与工程学院副教授. 主要研究方向为计算机视觉、图像/视频处理.
E-mail: wangchong@nbu.edu.cn