

基于三维卷积与自注意力机制的 视频行人重识别

庄珊娜^{1,2}, 王君帅^{1,2}, 白晶^{1,2}, 杜京瑾^{1,2}, 王正友^{1,2}

(1. 石家庄铁道大学 信息科学与技术学院, 石家庄 050043; 2. 河北省电磁环境效应与信息处理重点实验室, 石家庄 050043)

摘要: 针对视频行人重识别通常存在行人姿势多变、特征融合损失和遮挡干扰等问题, 本文提出了三维卷积与自注意力机制融合的网络模型。该模型首先通过三维卷积获取相邻帧之间的短期运动特征和局部细节特征; 其次提出多分支结构的跨尺度融合模块, 融合时间和语义维度的多阶段特征; 最后利用自注意力机制模块捕获长时信息。本文所提网络在 3 个经典公开数据集上实验, 并与其他方法进行对比, 实验结果表明, 该网络可有效利用视频长、短时信息, 获得较优性能。

关键词: 计算机应用; 视频行人重识别; 时序信息; 三维卷积; 自注意力机制

中图分类号: TP391.4 **文献标志码:** A **文章编号:** 1671-5497(2025)07-2409-09

DOI: 10.13229/j.cnki.jdxbgxb.20230881

Video-based person re-identification based on three-dimensional convolution and self-attention mechanism

ZHUANG Shan-na^{1,2}, WANG Jun-shuai^{1,2}, BAI Jing^{1,2}, DU Jing-jin^{1,2}, WANG Zheng-you^{1,2}

(1. School of Information Science and Technology, Shijiazhuang Tiedao University, Shijiazhuang 050043, China; 2. Hebei Provincial Key Laboratory of Electromagnetic Environmental Effects and Information Processing, Shijiazhuang 050043, China)

Abstract: To solve the problem of person posture change, feature fusion loss and blocking interference in video-based person re-identification, a fusion network model of three-dimensional convolution and self-attention mechanism is proposed in this paper. Firstly, short-term motion features and local detail features between adjacent frames are obtained through three-dimensional convolution. Secondly, a cross scale fusion module with multi-branch structure is proposed, multi-stage features of time and semantic dimensions are fused. Finally, long-term features are captured by a self-attention mechanism. Compared with other methods, this paper proposed network was tested on three commonly used public datasets, the experiments results demonstrate that the proposed network can efficiently utilized the long-term and short-

收稿日期: 2023-08-19.

基金项目: 国家自然科学基金项目(62341121); 河北省自然科学基金项目(F2023210001); 河北省教育厅青年基金项目(QN2022086, QN2025052); 河北省社会科学基金项目(HB19JY017).

作者简介: 庄珊娜(1985-), 女, 副教授, 博士. 研究方向: 计算机视觉及行人重识别. E-mail: zhuangshn@stdu.edu.cn

通信作者: 白晶(1990-), 女, 讲师, 博士. 研究方向: 信息处理及计算机视觉. E-mail: jingbai01@126.com

term information of videos so that advanced performances can be obtained.

Key words: computer application; video-based person re-identification; temporal information; three-dimensional convolution; self-attention mechanism

0 引言

行人重识别是智能视频监控领域的一项关键技术,旨在判别由不同视域摄像机在不同时间采集的行人视频/图像是否为同一目标行人^[1]。基于视频的行人重识别在智能安防、刑事侦查等领域有着广泛应用。来自监控视频的图像通常是模糊的,并且分辨率较低;不同的人身上可能会出现形状和穿着相似的情况;由于所获取的图像可能是在不同的时间拍摄的,因此行人的姿势可能不同;在现实生活中的视频监控场景中,还会出现遮挡等问题。在这些情况下,仅通过外表特征很难重识别不同的行人。

为解决上述问题,国内外学者进行了大量研究。Isobe 等^[2]提出了一个级联模块,有效地聚合片段级特征。Li 等^[3]提出了一个关系模块计算相关性向量,生成空间级和帧级的全局关系向量,帧级特征能够捕捉局部和全局信息。Hou 等^[4]利用注意力进行显著性特征擦除,以获得更丰富的空间特征。Tai 等^[5]提出了一种基于 Transformer 的网络,使其能够依赖较弱的监督信息学习目标各部分的特征。Carreira 等^[6]提出了膨胀三维卷积架构,该架构在时间维度上对来自 ImageNet 的预训练权重进行膨胀初始化模型权重。Li 等^[7]构建了一个多尺度三维(Multi-scale three-dimensional, M3D)卷积网络学习多尺度时间线索。

上述基于深度学习的方法对解决行人重识别中存在的遮挡、光照变化、相机角度、场景变化、行人姿态变化等问题有一定的效果,且其中一些方法通过注意力机制学习行人图像的显著性特征,但是这些方法缺乏对时间特征、全局特征和局部细节特征的统筹考虑。为此,本文提出一个三维(Three-dimensional, 3D)卷积与自注意力机制^[8]相融合的网络,该网络将所输入的行人视频序列分为 2 个子片段,一个子片段保留原始分辨率获取细节视觉线索,通过下采样处理另一子片段得到的低分辨率帧,并扩大感受野编码远距离空间上下文得到更广泛的特征^[9]。然后使用 3D 卷积和自注意力机制分别提取序列的短时信息和长时信息。本文所提网络通过结合短期和长期时间线索学习有判别性的时间表示,可增强模型面对视频中复杂多变环境的抗干扰能力,从而提高视频行人重识别的准确率。

1 本文方法

1.1 网络结构

本文构造了一种基于 3D 卷积和自注意力机制的视频行人重识别网络,网络结构如图 1 所示。该网络以 ResNet-50 作为基线网络,首先将输入的视频序列进行分段处理,利用不同的分辨率处理子片段。其次将 3D 卷积引入骨干网络

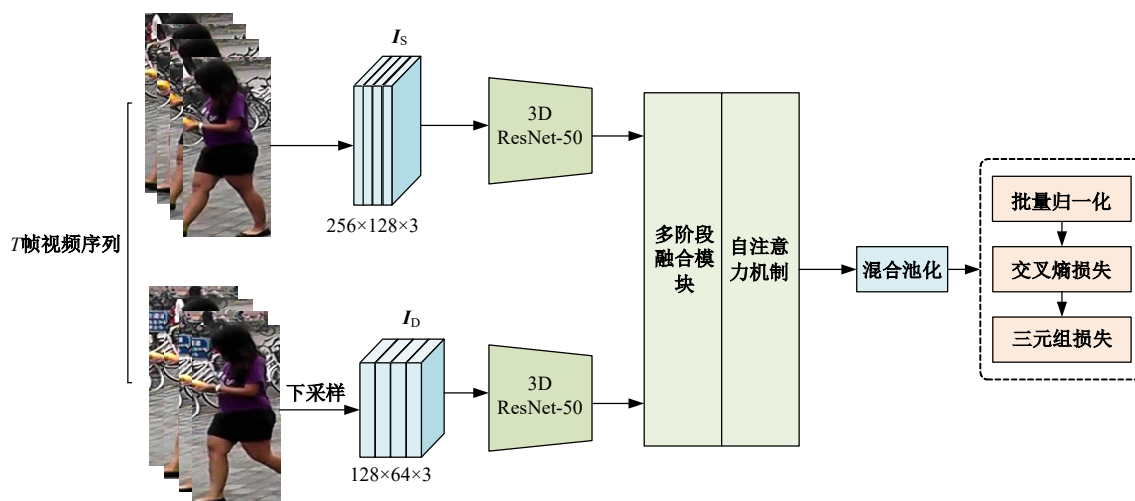


图 1 3D 卷积与自注意力机制融合网络结构

Fig. 1 Fusional network structure of 3D convolution and self-attention mechanism

ResNet-50中构成3D ResNet-50网络,用以获取片段的短时特征。再次将其输入自注意力机制模块中探索全局时间线索,得到视频序列的长时特征。最后通过多阶段融合模块对片段级特征进行融合,获得多尺度特征用于识别。

给定一个包含 T 帧的视频序列,本文将给定视频序列划分为 N 个子片段, $N=2$ 。将 2 个子片段的帧图像分别用不同的分辨率进行处理,其中, $I_s = \{I_t\}_{t=1}^{\frac{T}{2}}$ 以原始分辨率处理, $I_D = \{I_t\}_{t=\frac{T}{2}+1}^T$ 以原始分辨率的 $1/2$ 处理,可提供更大的感受野编码远距离空间上下文。首先利用 3D ResNet-50 网络提取每个视频帧片段的特征,表示为 $F = \{F_1, F_2, \dots, F_T\}$ 。其次通过本文提出的跨尺度融合模块将子片段特征连接起来,得到同时包含时间维度与语义维度的特征。最后将包含短时信息的特征输入自注意力机制模块中,探索不连续帧之间的上下文关系,捕获长时特征。

1.2 加入 3D 卷积的行人特征提取方法

3D 卷积是针对视频建模比较自然的方式,它具有同时获取行人序列的空间和时间维度特征信息的能力,能够有效增强行人的特征表达能力。

采用 ResNet-50 网络作为骨干网络,该网络的核心部分由 4 个卷积层组成,每层均由两种残差模块构成,其中第一层、第二层、第三层和第四层中的残差模块数分别为 3、4、6、3, ResNet-50 的残差块之间采用跳跃连接的方式,使得网络可以更好地处理较深层的特征。本文将 3D 卷积引入 ResNet-50 网络中构成 3D ResNet-50 网络,可同时提取视频帧的空间特征及短期时序信息,网络中的关键模块结构如图 2 所示。

图 2 中 C 为输出特征图的通道数, 3^3 为 $3 \times 3 \times 3$ 卷积核。ResNet-50 中的 bottleneck 模块由 3 个卷积层组成,卷积核的大小分别为 $1 \times 1 \times 1$ 、 $3 \times 3 \times 3$ 、 $1 \times 1 \times 1$ 。该结构可快速地将顶层的信息传递给 ReLU 之前的层,采用身份连接和零填充处理 ResNet-50 基本模块,以避免增加浅层网络的参数量。本文通过替换 ResNet-50 网络中的部分残差块,将 3D 卷积模块嵌入骨干网络中。具体为:在第二层中使用 2 个 3D 卷积替换第二层中的第一和第三个残差模块;在第三层中使用 3 个 3D 卷积替换第三层中的第一、第二和第三个残差模块。

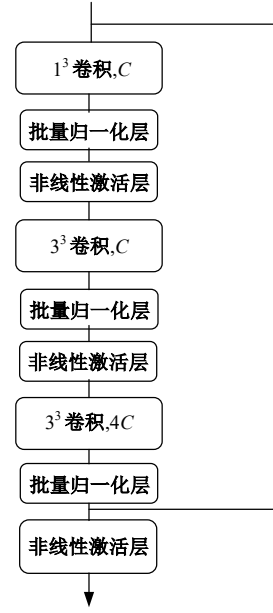


图 2 ResNet-50 中的 bottleneck 模块

Fig. 2 Bottleneck module in ResNet-50

1.3 自注意力机制

视频连续帧包含行人的运动信息,如行走姿势等,可用于辅助判断行人身份。对于时序建模,短期时间建模和长期时间建模都很重要,视频短时特征可描述人体细节信息,有助于区分外观相似的行人。但行人数据集采集环境较为复杂,仅靠短期时间线索不能解决问题,缺乏对全局信息的关注。因此需要捕捉全局时间线索,借用时间跨度比较大的帧的信息补全信息,缓解视频序列中的遮挡和噪声。长时信息可以概括视频的整体时序特征,减少低质量帧的影响。本文利用自注意力机制从全局角度捕获长时特征。在使用自注意力机制时,图像中的某个像素会和其他所有像素做相关性计算,即对于一个像素,充分考虑了整张图像。

图 3 给出了自注意力机制网络,自注意力机制输入为特征图 $X \in \mathbb{R}^{H \times W}$,其中 $\mathbb{R}$ 为实数集, H 和 W 分别为特征图的高度和宽度,通过不同的投影矩阵 $W^Q \in \mathbb{R}^{H \times H}$ 、 $W^K \in \mathbb{R}^{H \times H}$ 、 $W^V \in \mathbb{R}^{H \times H}$ 得到待训练参数矩阵 $Q = W^Q X$, $K = W^K X$, $V = W^V X$ 。对 Q 和每个 K 进行相似度计算得到权重,并归一化,将和 K 相应的 V 按归一化权重的注意力图进行加权求和,获得最终的输出特征。

1.4 多阶段融合模块

卷积神经网络是从浅层到深层的过程,先通过卷积层提取浅层特征,随着层数加深逐渐提取

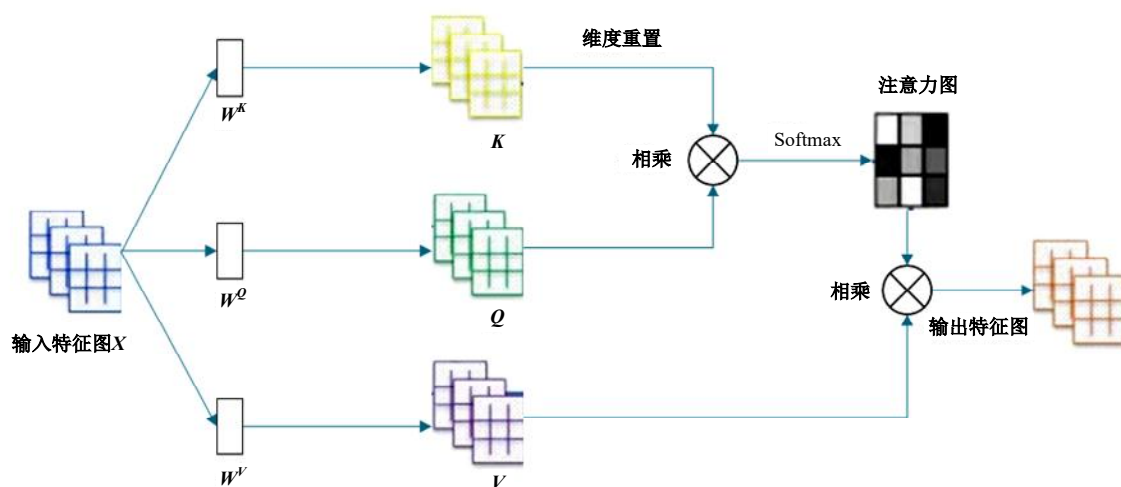


图 3 自注意力机制

Fig. 3 Self-attention mechanism

深层特征,可获取更多语义信息。浅层特征更关注图片的边缘和纹理等特征,深层特征能够学到更丰富的语义特征。本文提出一个多阶段融合模块(Multi-stage fusion module, MSFM)对不同阶段的特征进行融合,该模块在多个语义层面融合更丰富的帧特征,可有效减少传统单级融合造成的信息损失。

如图 4 所示,MSFM 包括两部分,分别是跨尺

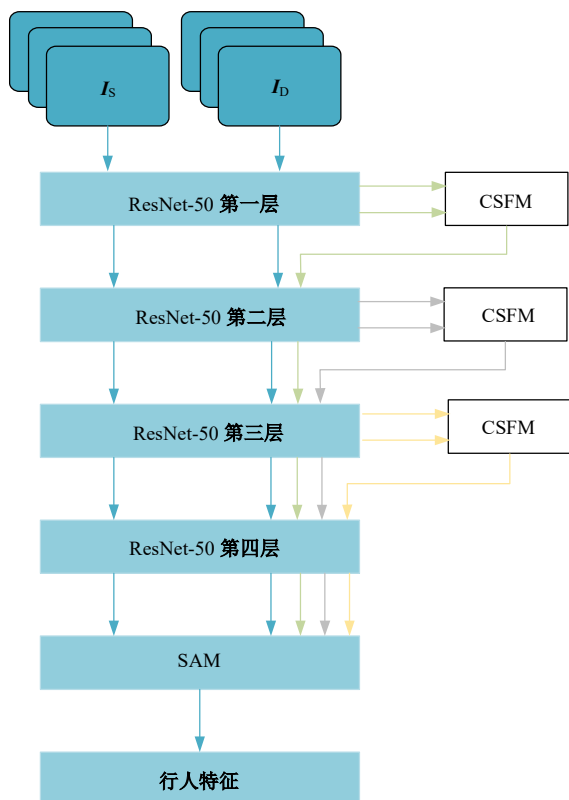


图 4 MSFM 结构示意图

Fig. 4 Schematic diagram of MSFM

度融合模块(Cross-scale fusion module, CSFM)和语义注意力模块(Semantic attention module, SAM)。下面对其进行详细介绍。

CSFM 主要作用于视频帧的特征,将低分辨率分支的特征传播到高分辨率分支,此时,高分辨率分支融合所有分支特征,可获得丰富的多粒度特征。CSFM 的结构示意图如图 5 所示。具体来说, $F_t \in \mathbb{R}^{H \times W \times C}$ 和 $F_{t+1} \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times C}$ 是 2 个分支同一阶段提取的特征图, H 、 W 和 C 分别为特征图的高度、宽度和通道数。CSFM 首先在 F_{t+1} 上进行上采样操作将其调整到与 F_t 相同的大小,然后将特征图 F_t 与 F_{t+1} 逐元素相加,得到融合后的特征。在语义维度上,将所有阶段的特征融合在一起,得到最终视频序列的特征。

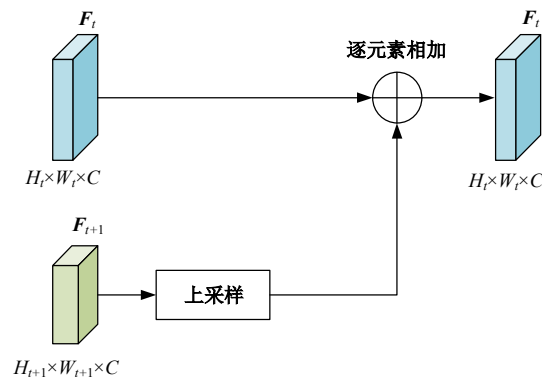


图 5 CSFM 结构示意图

Fig. 5 Schematic diagram of CSFM

本文使用 SAM 获取语义维度上的注意力权重。给定来自不同阶段的特征,通过 Softmax 得到每个阶段输出的重要性,然后求平均得到每个分支的注意力权重,最终多个语义阶段的特征按

注意力权重融合为视频级特征。

2 实 验

2.1 数据集

为验证本文网络模型的有效性,在 MARS^[10]、DukeMTMC-VideoReID^[11]和 iLIDS-VID^[12]数据集上进行实验验证本文所提方法的优越性。并且通过一系列消融实验,验证本文网络模型中各个模块对模型的有效性影响。

2.2 实验设置

本实验基于 Pytorch 框架,Ubuntu 64 位操作系统,显卡为 NVIDIA GeForce RTX 2080 Ti。在训练过程中,对于每个视频序列,以 4 帧的步幅随机采样 8 帧以形成一个视频片段。每一批包含 16 人,每人有 4 个视频片段。输入图片的统一高度为 256,宽度为 128,数据增强利用水平翻转和随机擦除^[13]。本实验中优化器采用 Adam^[14],权重衰减为 0.000 5,训练轮数为 150 轮,学习率初始化为 3.5×10^{-4} ,每 40 轮衰减因子为 0.1。在测试过程中,对于每个视频序列,首先将其拆分为几个 8 帧视频片段;其次通过提出的网络提取每个视频片段的特征,最终的视频特征是所有片段的平均表示;最后对计算查询和库特征之间的余弦距离进行检索。

2.3 实验分析

2.3.1 加入 3D 卷积的对比实验

本节主要通过对对比实验讨论在基线网络 ResNet-50 中添加 3D 卷积的最优嵌入方式,即在网络中添加 3D 卷积的位置和数量,以达到模型的准确率最佳。首先验证将 3D 卷积添加到骨干网络的哪一个卷积层中,能得到最好的实验结果。本次实验在 MARS 数据集上进行,将 3D 卷积加入 ResNet-50 各层中,实验过程中只用 3D 卷积替换每层中的第一个卷积,得到将 3D 卷积嵌入各卷积层中的实验结果,结果如表 1 所示。

根据表 1 结果可得以下结论:①在 ResNet-50 的不同位置中嵌入 3D 卷积得到的精度是不同的,当 3D 卷积只插入一层中时,在第三层中加入 3D 卷积的效果最好;②在 ResNet-50 网络的不同层中分别加入一个 3D 卷积所得结果是不同的。在本实验中,分别在第二层和第三层中嵌入一个 3D 卷积,该实验结果要优于单独在第二层或者第三层嵌入一个 3D 卷积。与基线网络相比,mAP 和

表 1 3D 卷积嵌入各层中的实验结果

Table 1 Results of 3D convolution embedding in each layer

方法	插入的层级	MARS	
		mAP	Rank-1
基线	—	81.2	86.6
	第一层	81.6	86.7
	第二层	81.8	86.8
	第三层	81.9	86.9
	第四层	81.7	81.7
	第二层+第三层	82.2	87.1

Rank-1 分别提高了 1.0 个百分点和 0.5 个百分点。

表 2 给出了 3D 卷积嵌入 ResNet-50 第二层中的实验结果。从表 2 中可看出,当用 3D 卷积替换第二层不同位置的卷积所得结果是不同的。并且用 3D 卷积替换 Conv3_2 的卷积时,准确率是最高的,用 3D 卷积替换多个位置的卷积所得结果也有差异,当第二层中的 Conv3_0 和 Conv3_2 残差块同时被 3D 卷积所替换时,实验结果最优,mAP 为 82.2%。

表 2 3D 卷积嵌入第二层中的实验结果

Table 2 Results of 3D convolutional embedding in layer 2

方法	模块数量	替换残差块的位置	MARS	
			mAP	Rank-1
基线	1	Conv3_0	81.6	86.7
		Conv3_1	81.4	86.7
		Conv3_2	82.0	86.8
		Conv3_3	81.5	86.6
	2	Conv3_0, Conv3_2	82.2	86.9
	3	Conv3_0, Conv3_2, Conv3_3	82.0	86.8

3D 卷积嵌入 ResNet-50 第三层中的实验结果如表 3 所示。由表 3 可得以下结论:①将一个 3D 卷积嵌入第三层中的不同位置的结果是有差异的,将 Conv4_4 残差块替换成 3D 卷积的结果是最优的;②将多个 3D 卷积嵌入第三层中的结果是不同的,当 Conv4_2 和 Conv4_4 位置的残差块分别被一个 3D 卷积替换时,网络的准确率较高;③当替换模块数量为 3 时,网络的准确率是最高的,mAP 为 83.1%,Rank-1 指标为 87.2%。但当替换模块数量为 4 时,实验结果有所下降。

将 3D 卷积同时嵌入 ResNet-50 第二、三层中的对比实验结果如表 4 所示。由表 4 可以看出,将在第二层和第三层中所得到的最优嵌入方式结合起来的准确率最高,与单独在第三层中嵌入 3D 卷

积相比,mAP提高了0.7个百分点,Rank-1提高了0.3个百分点。与单独在第二层中加入3D卷积相比,mAP和Rank-1分别提高了1.6个百分点和0.6个百分点。

表3 3D卷积嵌入第三层中的实验结果

Table 3 Results of 3D convolutional embedding in layer 3

方法	模块数量	替换残差块的位置	MARS	
			mAP	Rank-1
基线	1	Conv4_0	81.9	86.9
		Conv4_1	81.8	86.8
		Conv4_2	82.3	87.0
		Conv4_3	81.7	86.9
		Conv4_4	82.5	87.0
		Conv4_5	81.7	86.8
	2	Conv4_2, Conv4_4	82.7	87.1
	3	Conv4_0, Conv4_2, Conv4_4	83.1	87.2
	4	Conv4_0, Conv4_2, Conv4_4, Conv4_5	83.0	86.8

表4 3D卷积在两层中的结合对比实验

Table 4 Comparative experiment on the combination of 3D convolutions in two layers

方法	插入层级	替换残差块的位置	MARS	
			mAP	Rank-1
基线	第二层	Conv3_0, Conv3_2	82.2	86.9
	第三层	Conv4_0, Conv4_2, Conv4_4	83.1	87.2
	第二层	Conv3_0, Conv3_2	83.8	87.5
	第三层	Conv4_0, Conv4_2, Conv4_4	83.8	87.5

2.3.2 加入自注意力机制的对比实验

在2.3.1节中确定了3D卷积在ResNet-50网络中的最优嵌入方式,在此基础上进行加入自注意力机制的对比实验。本节在MARS和iLIDS-VID数据集上进行实验,结果如表5所示。

表5 自注意力机制对实验的影响

Table 5 Influence of self-attention mechanism on experiments

方法	MARS		iLIDS-VID	
	mAP	Rank-1	Rank-1	Rank-5
基线	81.2	86.6	84.0	95.0
基线+3D卷积	83.8	87.5	87.8	96.2
基线+自注意力	84.0	87.8	88.5	96.8
基线+3D卷积+自注意力	84.6	88.8	90.4	97.3

由表5可知,单独将自注意力机制加入到基线网络中,在MARS数据集上mAP和Rank-1分别为84.0%和87.8%。与基线网络相比,mAP和Rank-1分别提高了2.8个百分点和1.2个百分点。同时,将3D卷积和自注意力机制结合在一起

的结果是最优的,在MARS数据集上mAP和Rank-1分别为84.6%和88.8%。与基线网络相比,其关键指标mAP和Rank-1分别提高了3.4个百分点和2.2个百分点。

2.3.3 分支个数对实验结果的影响

本实验将视频序列分成多组片段,并对每组片段中的帧使用不同的分辨率进行处理。本文将不同分支对识别准确率的影响进行了比较,结果如表6所示。

表6 分支个数对实验的影响

Table 6 Influence of number of branches

分支1		分支2		MARS	
高	宽	高	宽	mAP/%	Rank-1/%
256	128	—	—	84.8	88.5
128	64	—	—	83.4	87.0
64	32	—	—	67.9	77.6
256	128	128	64	85.2	89.5
256	128	64	32	84.0	88.0
128	64	64	32	78.6	84.1

由表6可知:①在单分支的情况下,采用 256×128 分辨率训练网络可产生最好的结果,mAP和Rank-1分别为84.8%和88.5%。以 128×64 分辨率训练仍能提供合理精度,但当输入分辨率为 64×32 时,分辨率过低,性能严重下降;②在双分支结构中,当分辨率大小分别为 256×128 和 128×64 时,结果是最优的,mAP和Rank-1分别为85.2%和89.5%,与分辨率为 256×128 的单分支结构相比,mAP和Rank-1分别提高了0.4个百分点和1.0个百分点,Rank-1指标的提升效果显著。当分辨率大小分别为 256×128 和 64×32 时,模型性能显著下降,具有 64×32 输入分辨率的分支干扰网络参数的优化。

2.3.4 MSFM对实验结果的影响

本节评估MSFM对视频行人重识别准确率的影响,验证MSFM的有效性。MSFM包含CS-FM和SAM模块,将其加入基线网络进行消融实验,结果如表7所示。

由表7可知,在基线网络中加入MSFM可以提高模型准确率,基线网络是将分支1与分支2所得特征简单地连接起来。在MARS数据集上与基线网络对比,加入MSFM时mAP和Rank-1指标分别提高了0.6个百分点和0.6个百分点。当仅加入CSFM时,对比基线网络mAP和Rank-1指标分别提高了0.3个百分点和0.2个百分点。

表 7 MSFM 的消融实验

Table 7 Ablation experiment of MSFM %

方法	MARS	
	mAP	Rank-1
分支 1+分支 2	85.2	89.5
分支 1+分支 2+CSFM	85.5	89.7
分支 1+分支 2+SAM	85.6	89.9
分支 1+分支 2+MSFM	85.8	90.1

将 CSFM 和 SAM 结合在一起构成 MSFM 加入网络,实验结果达到最优。

2.4 方法对比

将本文方法与现有视频行人重识别方法进行比较。这些方法均为近 3 年所提出的,并使用 ResNet-50 作为其主干网络探索视频中行人的时空信息。对比结果见表 8,可以观察到,与其他比较方法相比,本文方法在 3 个数据集上都取得了显著的性能。

由表 8 可知,本文方法在 MARS 数据集上的 mAP 和 Rank-1 指标分别为 86.2% 和 90.7%,优于其他算法。相较于 TCLNet,本文方法在 iLIDS-

VID 数据集上 Rank-1 指标提高了 6.1 个百分点。在 MARS 数据集上 mAP 指标提高了 1.1 个百分点,这些显著的提升可归因于本文方法中双分支架构的设计。本文方法对短期和长期时间关系同时进行建模,所得识别性能优于 M3D 和 SSN3D 等仅对短期时间关系进行建模的方法。

由表 8 可知,本文方法在 DukeMTMC-Vid-eoReID 数据集上的 mAP 和 Rank-1 指标分别为 96.2% 和 97.0%,也取得了较好的识别结果。而 CSA-Net 对遮挡等低质量帧的处理,使得识别性能更优,因此对遮挡帧的处理也是本文下一步将继续改进的方向。

由表 8 可知,在 iLIDS-VID 数据集上,本文方法在 Rank-1 和 Rank-5 指标上的精度分别为 92.7% 和 98.9%,处于先进水平,低于 SITN 网络的 94% 和 100%。这是由于 SITN 网络在空间信息学习过程中时间信息起到了积极的指导作用,而 iLIDS-VID 数据集是冬天在机场采集的,行人外观信息相似,时间信息对空间信息的学习指导可充分发掘和利用细节信息。

表 8 与其他方法的对比

Table 8 Comparison with other methods %

方法	来源	MARS		DukeMTMC-VideoReID		iLIDS-VID	
		mAP	Rank-1	mAP	Rank-1	Rank-1	Rank-5
M3D ^[7]	IEEE2020	79.5	88.6	93.7	95.5	86.7	—
FGRA ^[15]	AAAI2020	81.2	87.3	—	—	88.0	96.7
VRSTC ^[16]	CVPR2020	82.3	88.5	93.5	95.0	83.4	—
STGCN ^[17]	CVPR2020	83.7	89.9	95.7	97.2	—	—
MG-RAFA ^[18]	CVPR2020	85.9	88.8	—	—	88.6	98.0
AFA ^[19]	ECCV2020	82.9	90.2	95.4	97.2	88.5	96.8
TCLNet ^[4]	ECCV2020	85.1	89.8	96.2	96.9	86.6	—
GRL ^[20]	CVPR2021	84.8	91.0	93.8	95.0	90.4	98.3
BiCnet-TKS ^[21]	CVPR2021	86.0	90.2	96.1	96.3	—	—
SSN3D ^[22]	AAAI2021	86.2	90.1	96.3	96.8	88.9	97.3
SINet ^[23]	CVPR2022	86.2	90.0	—	—	92.5	—
TE-AGC ^[24]	ACCV2022	85.8	90.7	96.3	97.2	—	—
GPNet ^[25]	Elsevier2023	86.0	90.5	96.3	97.3	88.5	98.5
SITN ^[26]	IEEE2023	85.1	89.2	96.0	97.2	94.0	100
CSA-Net ^[27]	ACM2023	84.5	90.4	96.7	97.7	90.0	98.3
本文	—	86.2	90.7	96.2	97.0	92.7	98.9

3 结束语

本文提出了一个基于 3D 卷积与自注意力机制的融合网络。长时信息可概括视频的整体时序

特征,减少低质量帧的影响,而 3D 卷积可捕获局部时间上下文信息,所得的特征包含了细节信息。该网络采用短时 3D 卷积和长时自注意力机制模块结合分别提取长短期时序特征进行互补,得到

完整的视频时序表示,有助于提高较大规模复杂场景下行人身份的识别准确率。此外,本文提出的 MSFM 将不同分支和不同阶段的特征相融合,得到一个同时包含了时间和语义维度的多阶段特征,在多个语义层面融合更丰富的帧特征,可有效减少传统单级融合造成的信息损失。对比实验及验证实验证明,本文所提网络能有效解决相机角度、场景变化、行人形状相近和遮挡等问题对视频行人重识别带来的干扰,提高视频行人重识别的准确率。

参考文献:

- [1] 于鹏, 朴燕. 基于多尺度特征的行人重识别属性提取新方法[J]. 吉林大学学报: 工学版, 2023, 53(4): 1155-1162.
Yu Peng, Piao Yan. New method for extracting person re-identification attributes based on multi-scale features[J]. Journal of Jilin University (Engineering and Technology Edition), 2023, 53(4): 1155-1162.
- [2] Isobe T, Han J, Zhuz F, et al. Intra-clip aggregation for video person re-identification[C]//Proceeding of the IEEE International Conference on Image Processing (ICIP). Piscataway, NJ: IEEE, 2020: 2336-2340.
- [3] Li X, Zhou W, Zhou Y, et al. Relation-guided spatialattention and temporal refinement for video-based person re-identification[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 11434-11441.
- [4] Hou R, Chang H, Ma B, et al. Temporal complementary learning for video person re-identification [C]//The 16th European Conference on Computer Vision, Glasgow, UK, 2020: 388-405.
- [5] Tai K S, Socher R, Manning C D. Improved semantic representations from tree-structured long short-term memory networks[C]//Proceeding of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Beijing, China, 2015: 1556-1566.
- [6] Carreira J, Zisserman A. Quo vadis, action recognition? A new model and the kinetics dataset[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Hawaii, USA, 2017: 6299-6308.
- [7] Li J, Zhang S, Huang T. Multi-scale 3d convolution network for video based person re-identification[C]//Proceedings of the AAAI Conference on Artificial Intelligence, Hawaii, USA, 2019, 33(1): 8618-8625.
- [8] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, Los Angeles, USA, 2017: 30-38.
- [9] 侯春萍, 杨庆元, 黄美艳, 等. 基于语义耦合和身份一致性的跨模态行人重识别方法[J]. 吉林大学学报: 工学版, 2022, 52(12): 2954-2963.
Hou Chun-ping, Yang Qing-yuan, Huang Mei-yan, et al. Cross-modality person re-identification based on semantic coupling and identity-consistence constraint [J]. Journal of Jilin University (Engineering and Technology Edition), 2022, 52(12): 2954-2963.
- [10] Zheng L, Bie Z, Sun Y, et al. Mars: a video benchmark for large-scale person re-identification[C]//The 14th European Conference on Computer Vision, Amsterdam, The Netherlands, 2016: 868-884.
- [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, 2018: 5177-5186.
- [12] Wang T, Gong S, Zhu X, et al. Person re-identification by video ranking[C]//The 13th European Conference on Computer Vision, Zurich, Switzerland, 2014: 688-703.
- [13] Zhong Z, Zheng L, Kang G, et al. Random erasing data augmentation[C]//Proceedings of the AAAI Conference on Artificial Intelligence, New York, USA, 2020, 34(7): 13001-13008.
- [14] Kingma D P, Ba J. Adam: a method for stochastic optimization[J/OL]. [2014-03-22]. <https://arxiv.org/pdf/1412.6980.pdf>.
- [15] Chen Z, Zhou Z, Huang J, et al. Frame-guided region-aligned representation for video person re-identification[C]//Proceedings of the AAAI Conference on Artificial Intelligence, New York, USA, 2020: 10591-10598.
- [16] Hou R, Ma B, Chang H, et al. Vrsrc: occlusion-free video person re-identification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Los Angeles, USA, 2019: 7183-7192.
- [17] Yang J, Zheng W S, Yang Q, et al. Spatial-temporal graph convolutional network for video-based person re-identification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2020: 3289-3299.

- [18] 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, USA, 2020: 10407-10416.
- [19] Chen G, Rao Y, Lu J, et al. Temporal coherence or temporal motion: Which is more critical for video-based person re-identification?[C]//The 16th European Conference on Computer Vision, Glasgow, UK, 2020: 660-676.
- [20] Liu X, Zhang P, Yu C, et al. Watching you: global-guided reciprocal learning for video-based person re-identification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Shenzhen, China, 2021: 13334-13343.
- [21] Hou R, Chang H, Ma B, et al. Learning efficient spatial-temporal representation for video person re-identification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, USA, 2021: 21-24.
- [22] Jiang X, Qiao Y, Yan J, et al. SSN3D: self-separated network to align parts for 3D convolution in video person re-identification[C]//Proceedings of the AAAI Conference on Artificial Intelligence, Online, 2021, 35(2): 1691-1699.
- [23] Bai S, Ma B, Chang H, et al. Salient-to-broad transition for video person re-identification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 2022: 7339-7348.
- [24] Ning J, Li F, Liu R, et al. Temporal extension topology learning for video-based person reidentification[C]//Proceedings of the Asian Conference on Computer Vision, Macao, China, 2022: 207-219.
- [25] Pan H, Chen Y, He Z. Multi-granularity graph pooling for video-based person re-identification[J]. Neural Networks, 2023, 160: 22-33.
- [26] Yang F, Li W, Liang B, et al. Spatiotemporal interaction transformer network for video-based person re-identification in internet of things[J]. IEEE Internet of Things Journal, 2023, 10(14): 12537-12547.
- [27] Wang K, Ding C, Pang J, et al. Context sensing attention network for video-based person re-identification[J]. ACM Transactions on Multimedia Computing, Communications and Applications, 2023, 19(4): 1-20.