基于Res2Net-Transformer的多尺度特征融合行人重识别算法

葛娟娟

（桂林理工大学 计算机科学与工程学院，广西 桂林 541006）

摘　要：为了能够准确地对行人图像进行识别，提出一种基于Res2Net-Transformer的多尺度特征融合行人重识别算法。该算法由全局特征提取模块、深度聚合模块和特征对齐模块组成，在全局特征提取模块中将Res2Net模块引入到ResNet50网络中，使网络提取到更细粒度的特征；通过多尺度深度聚合模块实现多尺度特征的循环聚合；利用特征对齐模块降低因特征不对齐对识别造成的影响。在Market1501、DukeMTMC-reID和MSMT17数据集上的实验结果表明，与现有方法相比，该方法具有更强的鲁棒性，在行人重识别上取得了更好效果。

关键词：行人重识别；图像识别；特征对齐；多尺度特征融合

DOI：10.11907/rjdk.241086　　开放科学（资源服务）标识码（OSID）：

中图分类号：TP391　　**文献标识码：**A　　**文章编号：**1672-7800(2025)003-0200-06

Person Re-identification Algorithm Based on Res2Net-Transformer for Multi-Scale Feature Fusion

GE Juanjuan

（*School of Computer Science and Engineering*, *Guilin University of Technology*, *Guilin* 541006, *China*）

Abstract: To accurately recognize pedestrian images , a person re-identification algorithm based on Res2Net-Transformer for multi-scale feature fusion is proposed. This method consists of a global feature extraction module, a deep aggregation module, and a feature alignment module. In the global feature extraction module, the Res2Net module is introduced into the ResNet50 network, enabling the network to extract more fine-grained features. The multi-scale deep aggregation module achieves the recursive aggregation of multi-scale features. The feature alignment module is used to mitigate the recognition impact caused by feature misalignment. Comparison with existing methods, this method approach demonstrates better robustness on Market1501、DukeMTMC-reID and MSMT17 dataset, it yields superior results in pedestrian re-identification.

Key Words: person re-identification; image recognition; feature alignment; multi-scale feature fusion

0　引言

行人重识别（Person Re-Identification，ReID）也称行人再识别，其目的是从给定的监控行人图像中，根据姿态、衣着等特征检索不同设备下是否存在相同的行人[1]。该技术在智能监控、自动驾驶、失踪人员搜寻等领域有着广阔的应用前景，亦是计算机视觉领域的热点话题。但由于监控摄像头分辨率不同，再加上光照变化、姿态改变及遮挡等问题的存在，给行人重识别方法的研究带来了诸多挑战。

随着深度学习的发展，基于卷积神经网络（Convolutional Neural Networks，CNN）的行人重识别方法取得了很大成功，如ResNet[2]、SeNet[3]、OSNet[4]等在行人重识别方面都有了广泛应用。但基于CNN的方法也存在一些问题，例如在对行人图像进行特征提取时，CNN会更多地关注局部特征，导致很多细节性的特征被忽略。为了解决以上问题，一些学者在行人重识别中引入注意力机制，例如Mixed High-order Attention Network（MHN）利用高阶注意力（High-Order Attention，HOA）模块建模，并利用复杂的高

收稿日期：2024-01-29
扫描二维码阅读全文：

作者简介：葛娟娟（1997-），女，桂林理工大学计算机科学与工程学院硕士研究生，研究方向为图像处理。

阶统计信息捕捉行人之间的细微差别[5]。Zhang等[6]提出一种全局关系感知注意模块(Relation-Aware Global Attention,RGA),通过对特征位置关系进行堆叠,使全局范围内的结构信息和局部外观信息更加紧凑,从而更好地进行注意力学习。虽然加入注意力机制能够在一定程度上改善CNN所存在的问题,但大多数注意力都无法捕捉到更大范围的上下文。如果将其放在全局范围内,很难实现对丰富结构化信息的利用。

近年来,随着Dosovitskiy等[7]将Transformer应用于计算机视觉领域,一些学者提出利用Transformer进行行人重识别。例如Wang等[8]提出的邻居Transformer网络NFormer,通过引入地标代理注意力和互惠邻居Softmax两个模块,在有效建模图像之间关系映射的低秩分解与特征空间少数地标的同时,还能够减轻不相关表示的干扰,从而抑制离群特征。然而,Transformer也存在一定不足,相较于CNN,其在位移、缩放、畸变不变性以及层次结构等方面表现较差。为此,也有学者提出将CNN与Transformer两种结构相结合进行行人重识别的方法。例如Zhang等[9]提出的HAT(Hierarchical Aggregation Transformer)模型将CNN与Transformer联合起来应用于行人重识别,引入基于Transformer的特征校准模块(Transformer-based Feature Calibration,TFC)插入到各个层次中,并提出深度监督聚合模块DSA(Deeply Supervised Aggregation)对来自各层次的网络特征递归地进行聚合,极大地提高了模型的识别能力。但是由于行人姿态的变化,再加上存在遮挡等问题,在进行特征对比时仅进行融合可能发生空间特征错位的情况。

针对以上问题,本文在文献[9]、[10]的启发下,提出基于Res2Net-Transformer的多尺度特征融合行人重识别算法。以ResNet50为主干网络,引入Res2Net模块以增强在更细粒度上的多个感受野,引入Transformer的特征校准模块(TFC),从全局视角探索信息并促进局部信息的融合,从而整合多尺度特征[11]。同时,引入特征对齐模块以减少空间错位导致的误差。最后在Market1501、Duke MTMC-reID、MSMT17 3个公开数据集上进行实验评估,结果证明,本文算法在行人重识别方面取得了较好效果。

1　网络整体结构

如图1所示,基于Res2Net-Transformer的多尺度特征融合的行人重识别方法主要由全局特征提取模块、深度聚合模块、特征对齐模块3部分组成。为了获得更细粒度的特征,本文在全局特征提取模块中以ResNet50作为骨干网络,以Res2Net模块替代ResNet50中的原始瓶颈模块,以行人图片作为输入进行全局行人特征的提取。在深度聚合模块中,通过引入TFC模块并将其插入Res2、Res3、Res4层中,以更好地对当前尺度特征的语义和细节信息进行整合,并利用监督聚合模块(Deeply Supervised Aggregation Model,DSAM)对聚合过程进行监督,从而实现对其多尺度特征从低级到高级聚合进行周期性的监督。在特征对齐模块中,将经过主干网络Res5提取的特征输入到特征对齐模块,然后利用水平方向上的全局池化对图像进行分割,以降低行人特征不对齐造成的影响。最后,为了对分类损失进行优化,使用平均池化、降维操作对图像维度进行转换。

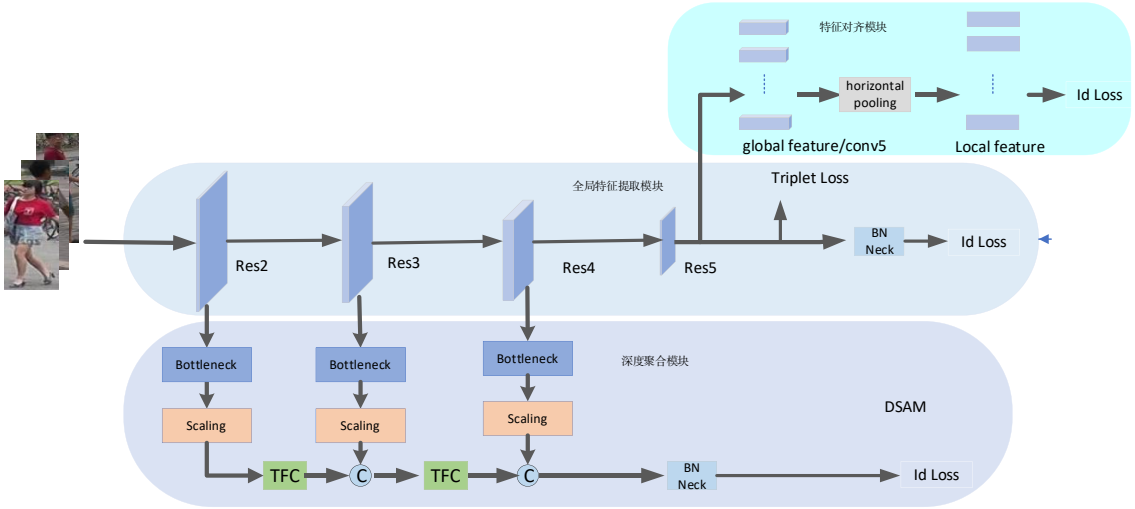

Fig. 1　Overall network architecture

图1　网络整体架构

1.1　全局特征提取模块

在行人重识别任务中,所提取到的特征信息越多,越能够对行人进行准确的识别。为了能够在更细粒度级别上获得多个可用的感受野,本文在骨干网络中引入Res2Net模块来替代原始的瓶颈模块。Res2Net模块结构如图2所示。与原始ResNet50瓶颈模块相比,Res2Net模块用一组较小的滤波器组对原始瓶颈模块n个通道的3×3卷积进行替换,同时以分层残差的方式将不同的滤波器组连接起来,使网络在拥有更强的多尺度特征提取能力的同时,还能够保持相似的计算负载。

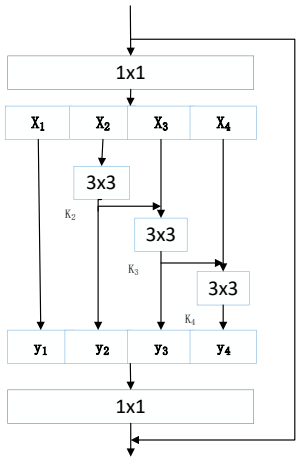

Fig. 2 Res2Net module

图 2 Res2Net 模块

具体来说，在进行 1 × 1 卷积后，输入的特征图会被均匀划分为 s 个不同的特征子集，用 x_i 表示，其中 $i \in \{1, 2, ..., s\}$。为了减少参数数量，除 x_1 外，每个 x_i 都有一个对应的 3 × 3 卷积，用 $K_i()$ 表示。对于被划分后的特征图，若用 y_i 表示各组输入特征对应的输出，则 y_i 可以表示为：

$$y_i = \begin{cases} x_i i = 1; \\ K_i(x_i) i = 2; \\ K_i(x_i + y_{i-1}) 2 < i \leqslant s. \end{cases} \tag{1}$$

即当 $i > 2$ 时，所得输出特征 y_i 为前一组输出特征 y_{i-1} 与下一组输入特征 x_i 相加后一起发送至下一组滤波器后经过 3 × 3 卷积后的结果。当 $i < s$ 时，该过程会一直持续，直到处理完为止。

当所有输入特征图被处理完获得输出特征后，所有组的输出特征将会被连接起来，发送至另一组 1 × 1 卷积中，将所有信息融合在一起。每个 3 × 3 卷积算子 $K_i()$ 可能会从所有分裂特征 $(x_j, j \leqslant i)$ 中接收特征信息，而每个特征子集 x_j 在经过 3 × 3 卷积后一般都能够输出比 x_j 更大的感受野。

本文用 Res2Net 模块替换原本在 ResNet50 中的原始瓶颈模块后，输入的图像特征经过网络后能够输出比原始更多数量与组合的感受野尺度，使图像能够在更细粒度级别上获得多个可用的感受野，更有利于全局和局部特征信息的提取。

1.2 深度聚合模块

融合多尺度特征信息能够在行人重识别过程中更准确地对行人进行识别，但是由于底层特征语义信息较少，简单的聚合结果会导致行人重识别的性能变差。因此，受 HAT 的启发，本文在深度聚合模块中引入 TFC 模块和 DSAM 监督聚合模块，通过串联的方式集成低级层次特征作为高级层次特征的全局先验，从而实现不同层次特征的聚合，并通过多粒度监督的方式完成对聚合过程的监督。

1.2.1 TFC 模块

为了增加不同层次特征之间的交互能力，在 TFC 模块中，利用 Transformer 对特征中的信息进行加强和抑制。该方式既能够捕获长距离的依赖关系，又能够从全局角度对多样化的信息进行关注，还能够在不同尺度特征交互过程中保持语义信息。TFC 模块具体结构如图 3 所示。

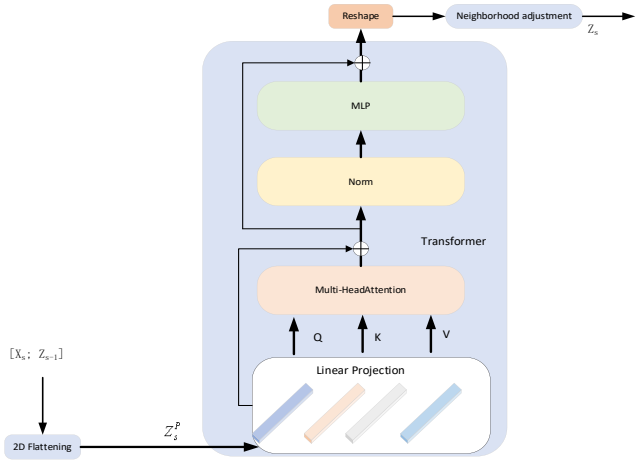

Fig. 3 TFC module

图 3 TFC 模块

为了便于后期对特征的集成和优化，首先应用残差块将 s 层提取到的特征 $X_s \in R^{C_s \times H \times W}$ 转换为紧凑的嵌入特征，其中 C_s、H、W 分别为当前特征图的通道数、高度和宽度。然后利用最大池化或双线性插值对分层特征进行调整，将其转换为相同的分辨率。

由于 TFC 模块需要接收令牌嵌入序列作为输入，将经过 TFC 模块上一层次的输出特征记为 $Z_{s-1} \in R^{C_{s-1} \times H \times W}$，则在进入 TFC 模块之前首先对输入的特征 $Z_s = [X_s; Z_{s-1}] \in R^{(C_s + C_{s-1}) \times H \times W}$ 进行处理，将其展平为 2D 的 patches Z_s^P，并通过对每个 patch 添加可学习的位置嵌入（Positional Embedding，PE）对空间信息进行整合。因此，每一个 patch 可以得到序列输入公式如下：

$$Z_s^p = \mathrm{PE}(Flattening(Z_s)) Z_s^P \in R^{(N+1) \times C_P} \tag{2}$$

其中，PE 表示对可学习的位置嵌入的添加，P 表示 patch 的大小，$N = H \times W / P^2$ 表示 patch 的数量，$C_p = C_s + C_{s-1} \times P^2$ 表示 patch 的通道数量，本文令 P 为 1。

网络获取到输入的序列后，将其输入到 Transformer 中进行处理。在 Transformer 中，特征将从头部送入，经由自注意力层生成 3 个不同的向量，并在不同特征之间建立可训练的关联关系。通过该方式可以在跨级别特征串联输入时实现在全局视图中多级信息的交互，从而更有效地捕捉到长距离的依赖关系。但是与 CNN 相比，Transformer 在位移、尺度、失真不变性以及层次结构等特性上仍存在不足。因此，为了使整个深度聚合模块能够联合 CNN 和 Transformer 的优点，在 Transformer 之后添加邻域调整模块对特征进行调整。其主要由批量归一化的卷积层堆栈组成，以保证能够将 ResNet50 的位移、尺度、失真不变性及提

取局部信息的能力与Transformer的能力相结合。所以，经过TFC模块的最终输出可以表示为：

$$Z_s = Conv(Reshape(Z_s^p)) \tag{3}$$

其中，*Reshape*表示重塑过程，将输出特征调整为与输入特征相同大小。

1.2.2　DSAM监督聚合模块

由于网络提取到的底层特征信息较少，若直接对不同层次进行聚合可能会导致模型的性能更差，因此本文引入DSAM监督聚合模块对TFC模块提取的特征进行聚合和监督。其跟随不同层次特征划分出的堆叠块进行迭代，对TFC模块所提取出的特征进行逐步聚合，并利用多粒度监督机制对聚合过程进行监督，从而减少直接对不同层次特征融合造成的对网络性能的限制。此外，为了对TFC模块交互过程中的语义信息提取能力进行增强，引入识别损失和三元组损失对分层聚合进行监督。具体公式如下：

$$DSAM(X_1, ..., X_n) = \begin{cases} X_1 & n=1 \\ TFC(Concat(DSAM(X_1, ..., X_{n-1}), X_n)) & \text{other} \end{cases} \tag{4}$$

最终，通过在深度聚合模块引入TFC模块和DSAM监督聚合模块，使模型能够实现在保留较高层次语义信息的同时，完成循环的对具有拥有更多细节信息和更少语义信息的深层特征的合并，在直接将较深层次和较浅层次连接的情况下完成多层次特征的聚合，从而更好地实现多尺度信息的融合。

1.3　特征对齐模块

在行人重识别中，由于所获取的行人图片中的信息比较杂乱，再加上行人的姿态发生变化，使得所查询的行人图像与原始行人图像往往在空间上会有一定区别，导致在对所提取到的特征进行比较时，由于空间出现一定的错位，无法对行人进行准确的识别。针对该问题，本文在网络中引入一个特征对齐模块，通过对所提取到的图片进行分割处理，以降低因为空间特征不对齐带来的误差。

如图1所示，特征对齐模块以主干网络Res5的输出特征作为分支的输入。若将Res5的输出特征记为$X \in R^{C \times H \times W}$，其中$C$表示通道数，$H$表示高度，$W$表示宽度。在特征对齐模块中，首先对所获取到的特征进行水平方向上的全局池化，并利用1×1卷积将原始的通道数C减少到c，从而得到H个代表图像水平部分的局部特征表示。特征分支的最终输出为：

$$\{y_1, y_2, ..., y_H\} = conv(HP(X)) \tag{5}$$

其中，HP代表水平方向上的全局池化。

通过分割，在进行最后的特征对齐的比较时，不再是对整个图像特征进行对齐的比较，而是对分割后的每H个局部特征进行比较。该方式可使对比的特征信息更加精细，对比结果也更加准确，从而在一定程度上降低了因特征不对齐带来的误差。

1.4　损失函数

为了学习到更有辨别性的特征，本文联合使用标签平滑损失和三元组损失对网络进行训练。其中，标签平滑损失用于对主干网络和Transformer网络训练过程进行监督。识别损失的计算公式如下：

$$L_{id} = -\sum_{i=1}^{N} q_i \log(p_i) \begin{cases} q_i = \dfrac{\varepsilon}{N}, y \neq i \\ q_i = 1 - \dfrac{N-1}{N}\varepsilon, y = i \end{cases} \tag{6}$$

其中，N是训练集的类别数，p_i是输出的行人图像属于第i类的预测概率，y是行人的真实标签信息，ε是用于提高泛化能力的一个常数。

三元组损失用于对不同的样本图像进行区分，从而使类内距离小于类间距离。若$d_{a,p}$、$d_{a,n}$分别表示正负样本对的相对距离，则具体公式如下：

$$L_{tri} = (d_{a,p} - d_{a,n} + \alpha)_+ \tag{7}$$

其中，$(\bullet)_+$表示$\max(\bullet, 0)$，α表示函数阈值参数。

综上所述，模型的总体损失函数为：

$$L_{reid} = L_{id} + L_{tri} + \sum_{j=1}^{n_{al}} (L_{al}) \tag{8}$$

其中，n_{al}为级联块数量，L_{al}由识别损失和三元组损失组成。

2　实验与分析

2.1　数据集及评价标准

为了验证本文方法的有效性，选用3个公开的数据集对实验进行评估。其中，Market1501数据集由6个摄像头采集，DukeMTMC-reID数据集是从8个摄像头采集的DukeMTMC数据集中截取得到的，MSMT17数据集由15个摄像头采集。具体如表1所示。

Table 1　Dataset introduction

表1　数据集介绍

数据集	训练集		测试集		总计	
	行人	图像	行人	图像	行人	图像
Market1501	751	12 936	750	19 372	1 501	32 668
DukeMTMC-reID	702	16 522	702	19 889	1 404	36 411
MSMT17	3 060	32 621	3 060	93 820	4 101	126 441

实验使用累积匹配特征（Cumulative Match Characteristic，CMC）中的Rank-k和平均精度（mean Average Precision，mAP）作为评价指标进行行人重识别模型性能评估。其中，Rank-k是前k个图像中出现所查询图像的概率，本文使用Rank-1作为评价指标。mAP是对所有类别平均精度进行综合加权平均的值。mAP的值越高，说明模型检测越准确。

2.2　实验设置

本文实验基于深度学习框架Pytorch，使用64位Ubuntu 20.04操作系统、RTX 3090的显卡对网络进行训练和评

估。在将图像输入网络前，将输入图像的大小调整为 256 × 128，并且利用随机翻转、随机裁剪和随机擦除等方式对训练集中的每张图像进行随机数据增强。之后，将 batchsize 大小设为 64，采用 Adam 作为网络优化器进行网络训练。

2.3 实验结果对比与分析

为验证本文算法的有效性，本文在 3 个数据集上进行实验，并将其与近年来的先进方法进行比较。由表 2 可见，在 Market1501 数据集上，本文方法在 mAP 与 Rank-1 上分别达到了 90.4% 和 96.2%，比其他方法最好的结果提升了 0.1% 和 0.2%。在 DukeMTMC-reID 数据集上，mAP 和 Rank-1 分别达到了 82.6% 和 92.3%，比其他方法最好的结果提升了 1.2% 和 1.3%。因此，在 Market1501 和 DukeMTMC-reID 数据集上，本文方法能够提取到鲁棒性更强的行人特征，相比其他方法具有更好的性能。

Table 2 Comparison results of different methods on Market1501 and DukeMTMC-reID

表 2 在 Market1501 与 DukeMTMC-reID 上不同方法对比结果

Methods	Market1501		DukeMTMC-reID	
	mAP	Rank-1	mAP	Rank-1
HA-CNN[19]	75.7	91.2	63.8	80.5
LANet[22]	83.1	94.4	73.4	87.1
PCB+RPP[10]	81.6	93.8	69.2	83.3
OSNet[4]	84.9	94.8	73.5	88.6
MHN[5]	85.0	95.1	77.2	89.1
CDNet[15]	86.0	95.1	76.8	88.6
CACE-Net[16]	**90.3**	**96.0**	81.3	90.9
PAT[20]	88.0	95.4	78.2	88.8
ABD-Net[17]	88.3	95.6	78.6	89.0
SCSN[18]	88.5	95.7	79.0	**91.0**
DGNet[21]	86.0	94.8	74.8	86.6
HAT[9]	89.8	95.8	**81.4**	90.4
本文方法	**90.4**	**96.2**	**82.6**	**92.3**

由表 3 可见，在 MSMT17 数据集上，本文方法在 mAP 与 Rank-1 上分别达到了 63.1% 和 83.8%，在 Rank-1 上与其他方法最好的结果持平，但在 mAP 上相比 SCSN 的结果提升了 4.6%，表明在 MSMT17 数据集上，本文方法在行人检测过程中能够取得更准确的结果。

通过以上对比可以看出，本文方法无论是在行人检测准确性方面，还是在提取到的行人特征鲁棒性方面都具有更优异的表现。

Table 3 Comparison results of different methods on MSMT17

表 3 在 MSMT17 上不同方法对比结果

Methods	MSMT17	
	mAP	Rank-1
LANet[22]	46.8	75.5
PCB+RPP[10]	40.4	68.2
OSNet[4]	52.9	78.7
CDNet[15]	54.7	78.9
CACE-Net[16]	62.0	83.5
ABD-Net[17]	60.8	82.3
SCSN[18]	58.5	**83.8**
DGNet[21]	52.3	77.2
HAT[9]	61.2	82.3
本文方法	**63.1**	**83.8**

2.4 消融实验

为了验证各模块的有效性，本文在 Market1501 数据集和 DukeMTMC-reID 数据集上分别进行了消融实验。具体来说，本文以原始的主干网络 ResNet50 和在主干网络的 Res2、Res3、Res4 引入深度聚合模块后的网络作为基线（Baseline），分别在基线的基础上增加 Res2Net 模块、特征对齐模块，通过消融实验验证每个模块对行人重识别结果的影响。

从表 4 可以看出，在 Market1501 数据集中，与 Baseline 相比，添加特征对齐模块后 mAP 与 Rank-1 分别提升了 0.2% 和 0.3%，添加 Res2Net 模块后 mAP 与 Rank-1 分别提升了 0.5% 和 0.4%。两个模块共同作用时，mAP 与 Rank-1 分别提升了 0.7%。

Table 4 Comparison of ablation experiment results on Market1501

表 4 在 Market1501 上消融实验结果比较

方法	mAP	Rank-1
Baseline	89.7	95.5
+特征对齐	89.9	95.8
+Res2Net	90.2	95.9
本文方法（+Res2Net 与特征对齐）	**90.4**	**96.2**

从表 5 可以看出，在 DukeMTMC-reID 数据集中，与 Baseline 相比，添加特征对齐模块后 mAP 与 Rank-1 分别提升了 0.3% 和 0.2%，添加 Res2Net 模块后 mAP 与 Rank-1 分别提升了 1.3% 和 0.9%。两个模块共同作用时，mAP 与 Rank-1 分别提升了 1.4% 和 2.0%。

Table 5 Comparison of ablation experiment results on DukeMTMC-reID

表 5 在 DukeMTMC-reID 上消融实验结果比较

方法	mAP	Rank-1
Baseline	81.2	90.3
+特征对齐	81.5	90.5
+Res2Net	82.5	91.2
本文方法（+Res2Net 与特征对齐）	**82.6**	**92.3**

通过在上述两个公开数据集上的消融实验结果可以看出，引入深度聚合模块的 Baseline 能够很好地对不同层次的多尺度特征进行融合。添加特征对齐模块能够在一定程度上降低原始网络中空间特征不对齐带来的影响，使模型检测的准确率得到一定提升，而添加 Res2Net 模块后能够在更细粒度的方向上提取到更多行人特征，更好地帮助网络对行人进行识别。因此，本文算法能够更准确地对行人进行识别，从而有效提升行人重识别的性能。

3 结语

本文提出一种基于 Res2Net-Transformer 的多尺度特征融合行人重识别算法，添加 Res2Net 模块对 ResNet50 网络进行改进，使其能够提取到更细粒度的特征，添加 TFC

模块可实现对多尺度特征的融合,同时构建特征对齐模块来降低特征不对齐带来的影响,最后使用标签平滑损失和三元组损失对网络进行联合训练。在Market1501、Duke MTMC-reID和MSMT17数据集上的实验对比结果证明,本文方法能够在行人重识别上取得较好性能。未来将进一步对模型性能进行优化,同时探索更高效的特征提取方法以增强跨领域应用的适应性。

参考文献:

[1] ZOU G F, FU G X, GAO M L, et al. Research progress on metric learning methods in pedestrian re-identification [J]. Control and Decision, 2021, 36(7): 1547-1557.
邹国锋,傅桂霞,高明亮,等. 行人重识别中度量学习方法研究进展[J]. 控制与决策, 2021, 36(7):1547-1557.

[2] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]// IEEE Conference on Computer Vision and Pattern Recognition,2016:770-778.

[3] HU J, SHEN L, SAMUEL A, et al. Squeeze-and-excitation networks [C]// IEEE Conference on Computer Vision and Pattern Recognition, 2018:7132-7141.

[4] ZHOU K, YANG Y, CAVALLARO A, et al. Omni-scale feature learning for person re-identification[C]//IEEE International Conference on Computer Vision,2019:3701-3711.

[5] CHEN B H, DENG W H, HU J. Mixed high-order attention network for person re-identification[C]//IEEE International Conference on Computer Vision,2019:371-381.

[6] ZHANG Z Z, LAN C L, ZENG W J, et al. Relation-aware global attention for person re-identification[C]//IEEE Conference on Computer Vision and Pattern Recognition,2020:3183-3192.

[7] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: transformers for image recognition at scale[DB/OL]. https://arxiv.org/abs/2010.11929.

[8] WANG H, SHEN Y, LIU Y, et al. NFormer: robust person re-identification with neighbor transformer[C]//IEEE Conference on Computer Vision and Pattern Recognition,2022:7287-7297.

[9] ZHANG G W, ZHANG P P, QI J Q, et al. HAT: hierarchical aggregation transformers for person re-identification[C]//Proceedings of the 29th ACM International Conference on Multimedia,2021:516-525.

[10] SUN Y, ZHENG L, YANG Y, et al. Beyond part models: person retrieval with refined part pooling(and a strong convolutional baseline)[C]//Proceedings of the European Conference on Computer Vision, 2018: 480-496.

[11] GAO S H, CHENG M M, ZHAO K, et al. Res2Net: a new multi-scale backbone architecture[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2021,43(2):652-662.

[12] CHEN Y, KUANG C. A multi-scale learning pedestrian re identification method based on CNN and Transformer [J]. Journal of Electronics and Information Technology, 2023, 45(6):2256-2263.
陈莹,匡澄. 基于CNN和Transformer多尺度学习行人重识别方法[J]. 电子与信息学报,2023,45(6):2256-2263.

[13] LUO H, JIANG W, GU Y Z, et al. A strong baseline and batch normalization neck for deep person re-identification[J]. IEEE Transactions on Multimedi,2020,22(10):2597-2609.

[14] ZHANG X, LUO H, FAN X, et al. AlignedReID: surpassing human-level performance in person re-identification[DB/OL]. https://arxiv.org/abs/1711.08184.

[15] LI H J, WU G J, ZHENG W S. Combined depth space based architecture search for person re-identification[C]//IEEE Conference on Computer Vision and Pattern Recognition,2021:6725-6734.

[16] YU F F, JIANG X Y, GONG Y F, et al. Devil's in the details: aligning visual clues for conditional embedding in person re-identification[DB/OL]. https://arxiv.org/abs/2009.05250.

[17] CHEN T L, DING S J, XIE J Y, et al. ABD-Net: attentive but diverse person re-identification[C]//IEEE International Conference on Computer Vision,2019:8350-8360.

[18] CHEN X S, FU C M, ZHAO Y, et al. Salience-guided cascaded suppression network for person re-identification[C]// IEEE Conference on Computer Vision and Pattern Recognition,2020:3297-3307.

[19] LI W, ZHU X T, GONG S G, et al. Harmonious attention network for person re-identification[C]//IEEE Conference on Computer Vision and Pattern Recognition,2018:2285-2294.

[20] LI Y L, HE J F, ZHANG T Z, et al. Diverse part discovery: occluded person re-identification with part-aware transformer[C]//IEEE Conference on Computer Vision and Pattern Recognition,2021:2897-2906.

[21] ZHENGZ D, YANG X D, YU Z D, et al. Joint discriminative and generative learning for person re-identification[C]//IEEE Conference on Computer Vision and Pattern Recognition,2019:2133-2142.

[22] HOU R B, MA B P, CHANG H, et al. Interaction-and-aggregation network for person re-identification[C]// IEEE Conference on Computer Vision and Pattern Recognition,2019:9309-9318.

(责任编辑:黄　健)