DOI: 10.16452/j.cnki.sdkjzk.2025.01.010 文章编号:1672-3767(2025)01-0110-09

基于 CNN-Transformer 和注意力金字塔的行人重识别方法研究

徐 岩[1],刘香兰[1,2],潘旭光[1],李 芳[1],赵海燕[1]
(1.山东科技大学 电子信息工程学院,山东 青岛 266590;
2.中国联合网络通信有限公司日照市分公司,山东 日照 276800)

摘 要:针对行人重识别技术中难以捕捉不同尺度特征图的显著区域,以及在非重叠摄像机中将多尺度特征汇总到全局视图中仍存在挑战的问题,提出一种基于 CNN-Transformer 和注意力金字塔的行人重识别方法。首先,引入基于 Transformer 的特征校准模块和深度监督聚合方法组成 TFCNet,从全局角度以循环自适应的方式将骨干网络的各层级不同尺度的特征进行聚合。然后,设计一种串行融合注意力模块,在计算时能够结合通道和空间的信息交互。同时,引入注意力金字塔,设计一种多尺度串行融合注意力金字塔结构,采用由粗到细的金字塔方法学习到更多不同尺度特征图的显著区域,提取更多有识别性的行人特征。最后,采用多重损失函数对网络进行总体优化,并在三个主流数据集上进行实验验证,证明了所提方法的有效性。

关键词:行人重识别; Transformer; 特征聚合; 注意力金字塔
中图分类号:TP391 **文献标志码**:A

Research on person re-identification method based on CNN-Transformer and attention pyramid

XU Yan[1], LIU Xianglan[1,2], PAN Xuguang[1], LI Fang[1], ZHAO Haiyan[1]
(1. College of Electronic and Information Engineering, Shandong University of Science and Technology, Qingdao 266590, China; 2. China Unicom Rizhao Branch, Rizhao 276800, China)

Abstract: A person re-identification method based on convolutional neural networks (CNN)-Transformer and attention pyramid was proposed to address the difficulty in capturing significant areas of different scale feature maps and the challenge of aggregating multi-scale features into a global view in non-overlapping cameras. The feature calibration module based on Transformer and deep supervision aggregation were introduced to form TFCNet to aggregate the features of different scales at all levels of the backbone network in a circular adaptive way from a global perspective. A serial fusion attention module that could combine channel and spatial information interaction during computation was designed. By introducing an attention pyramid, a multi-scale serial fusion attention pyramid structure was designed as well. More prominent regions of different scale feature maps were learned and more recognizable pedestrian features were extracted by using a coarse-to-fine pyramid method. Finally, multiple loss function was used to optimize the network. Experimental validation on three mainstream datasets demonstrates the effectiveness of the proposed method.

Key words: person re-identification; Transformer; feature aggregation; attention pyramid

收稿日期:2023-05-18
基金项目:山东省研究生教育优质课程项目(SDYKC19083);山东省研究生教育联合培养基地项目(SDYJD18027);海信集团海信研究院资助项目(SKDHKQ20240464)

作者简介:徐 岩(1970—),男,山东汶上人,教授,博士生导师,研究方向为人工智能与模式识别、图像识别与信号处理等.
刘香兰(1997—),女,山东海阳人,硕士研究生,研究方向为机器学习、深度学习,本文通信作者.
E-mail:17663350120@163.com

行人重识别(re-identification,ReID)是指在无重叠视野区域的摄像头下,对目标行人进行跨摄像头、跨场景下的检索和匹配。近年来,随着人工智能技术的进步,行人重识别技术水平取得显著提升。但由于图像在拍摄中容易受到不确定因素的影响,使得有效的特征信息在检测时容易被抑制,在识别过程中很难通过有区别性的特征找到目标行人。

早期基于传统方法的行人重识别研究中各个环节都是相互独立的,且仅适用于小数据集。随着深度学习的发展,将卷积神经网络(convolutional neural networks,CNN)应用在行人重识别任务中,能够从行人图片中提取丰富的粒度特征表示,但 CNN 模型擅长提取局部特征,在获取全局特征方面缺少建立远程依赖关系的能力。2017 年,Transformer 在自然语言处理(natural language processing,NLP)领域崭露头角并取得成功。2020 年,Dosovitskiy 等[1]提出用于图像识别任务的 Vision Transformer(ViT),不依赖 CNN 结构,能够学习到完整的全局特征表示,但受预训练数据量需求的限制,ViT 网络的参数量通常比较"重",往往会忽视局部特征细节,其结构缺乏诸如移位、尺度变化、层次结构等理想特性。

为了汲取 CNN 的轻量与 Transformer 的表征全局特征这两种优点,一些 CNN-Transformer 的混合模型开始在 ReID 中流行起来。2021 年,Peng 等[2]提出一种混合模型结构 Conformer,采用卷积和自注意力机制并行,并依靠特征耦合单元(feature coupling unit,FCU),以交互式最大限度地融合不同分辨率下的局部信息和全局信息。同年,基于 Transformer 的特征校准(transformer-based feature calibration,TFC)模块和深度监督聚合(deeply supervised aggregation,DSA)方法组成的新框架 HAT[3]被提出,首次在基于图像的行人重识别中采用了 CNN 和 Transformer 的结构。Guo 等[4]设计了一种串行的混合模型(convolutional neural networks meet vision transformers,CMT),在基于 Transformer 的结构中引入卷积操作,提出由 depth-wise 卷积和自注意力机制组成的 CMT 模块,可以同时捕获局部与全局的关系,在准确性、参数量以及计算复杂度方面获得了更好的权衡。Xu 等[5]基于注意力及残差思想提出残差注意力网络(residual attention network,RANet),对马铃薯叶部病害进行识别并验证了其有效性。

很多研究者探究了卷积神经网络各层级多尺度特征的有效性。例如,Chen 等[6]提出了显著特征提取单元(salient feature extraction unit,SFE Unit)用于抑制最显著的特征,挖掘潜在的特征,通过级联的方式逐级将挖掘的特征融合到最终的特征表示中。Liu 等[7]利用基于注意力的结构自适应地融合多级特征。

在行人重识别技术中,捕捉不同尺度和不同粒度特征图的显著区域存在困难,同时在非重叠摄像机中将多尺度特征整合到全局视图中也面临挑战,本研究提出一种基于 CNN-Transformer 和注意力金字塔的行人重识别网络(CNN-Transformer and attention pyramid network,CTAAPNet)。基于 Transformer 的特征校准模块和深度监督聚合(deeply supervised aggregation,DSA)方法构建(transformer-based feature calibration network,TFCNet),从全局角度出发,将骨干网络中各层级不同尺度的特征以循环自适应的方式予以聚合。设计了多尺度串行融合注意力金字塔网络,用来提取更多具有区别性的行人特征;同时,采用多重损失函数,总体优化了网络模型。在 Market-1501 数据集中,平均精度均值和 Rank-1 精度分别达到了 91.00%和 96.31%,与近年来其他一些主流网络的实验对比表明,所设计的方法表现出良好的鉴别力、鲁棒性和泛化能力,提出的优化方案对网络模型性能具有明显提升。

1　基于 CNN-Transformer 和注意力金字塔的行人重识别网络

1.1　网络结构及流程

本研究提出的基于 CNN-Transformer 和注意力金字塔的行人重识别网络(CTAAPNet),主要包括多尺度串行融合注意力金字塔网络(serial fusion attention pyramid network,SFAPNet)、深度监督聚合(DSA)方法和基于 Transformer 的特征校准模块(transformer-based feature calibration network,TFCNet),CTAAPNet 网络结构如图 1 所示。

模型训练阶段,输入的行人图像的大小为 256×128,通过随机擦除、随机水平翻转和随机裁剪等方法进行数据增强。首先,在卷积层 Conv3 和 Conv4 依次加入 SFAPNet 来强化模型对区别性特征的学习。然后,在 TFCNet 中将 TFC 模块插入到卷积层 Conv2、Conv3 和 Conv4 中,为下一层级特征生成全局先验。采用

深度监督聚合(DSA)方法,从浅层到深层逐步聚合细化特征表示,缓解浅层特征中语义信息较少的问题,并使用辅助损失监督层级的聚合。最后,结合 BN Neck[8] 对特征进行归一化处理。在全连接层(fully-connected layer,FC)前添加一个 BN(batch normalization)层,将 BN 层之前的特征记为 f_1,BN 层之后的归一化特征记为 f_2。在训练阶段,分别使用 f_1 和 f_2 计算三元组损失(L_{Triplet})和 ID 损失(L_{ID})并使之收敛。

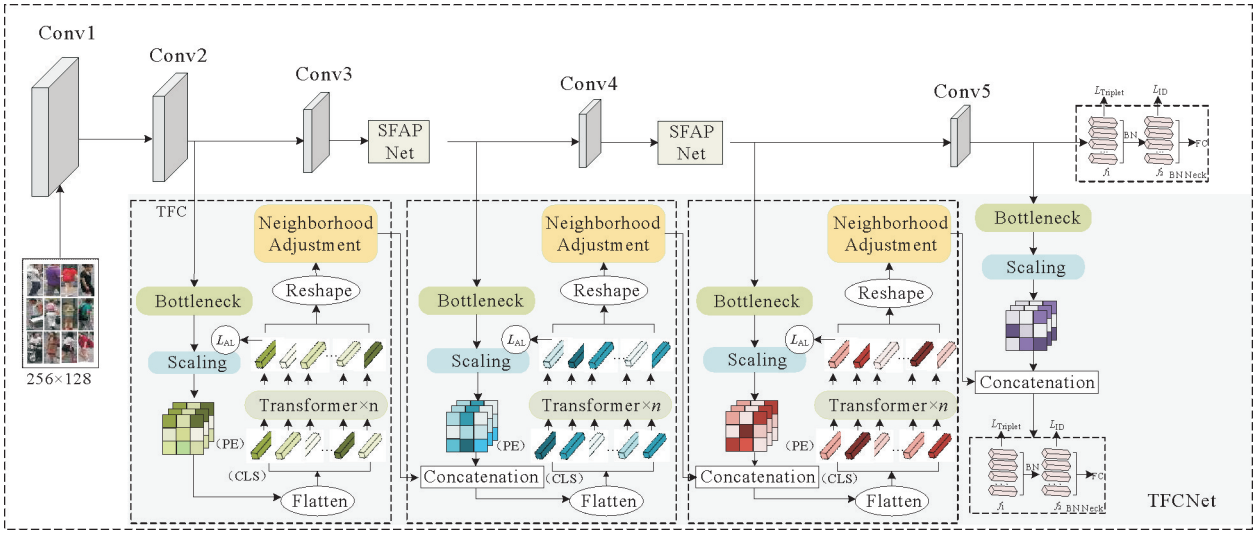

图 1 基于 CNN-Transformer 和注意力金字塔的行人重识别网络

Fig. 1 Person re-identification network based on CNN-Transformer and attention pyramid

利用高分辨率相机技术(high resolation camera,HiResCAM)对 ResNet50-IBN-a[9] 的后 4 个卷积层分别进行可视化处理,可视化结果如图 2 所示,可以看出,在卷积层 Conv3 和 Conv4 后包含更丰富的区别性特征,能够区分不同尺度的特征图中判别性较高的区域,更适用于模型的训练学习。

1.2 串行融合注意力模块

注意力机制是模仿人的一种思维模式,在训练时能重点关注某些信息[10]。压缩-激励模块(squeeze-excitation module,SE)可以从全局范围内捕获相关信息,而不受顺序方式的限制[11]。三重注意力模块实现了用近乎为"零"参数的注意力机制进行通道和空间的注意力对齐,弥补了许多注意力机制在空间和通道缺少交互机制的缺陷。

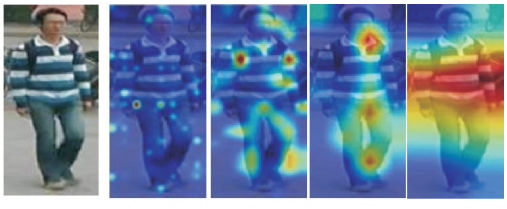

图 2 ResNet50-IBN-a 四个残差块的 HiResCAM 热力图

Fig. 2 HiResCAM heat map of the four ResNet50-IBN-a residual blocks

本研究融合 SE 和三重注意力模块,设计了串行融合注意力模块(serial fusion attention module,SFAM),如图 3 所示。首先 SE 模块的实现主要有两个步骤。

1) 压缩(Squeeze)。为了将全局空间信息压缩到通道描述符中,使 $\boldsymbol{X}$ 输出的每个单元能够利用该区域的上下文信息,原始特征图 $\boldsymbol{X}$ 经过全局平均池化(global average pooling,GAP),将每个通道的二维特征 $H\times W$ 压缩降维为 1×1,每个通道都有一个数值表示,数值的大小代表该通道的重要程度,且数值包含该通道的全局感受野,即特征图由 $\boldsymbol{X}(H\times W\times C)$ 变为向量 $\boldsymbol{Z}(1\times 1\times C)$,因此 $\boldsymbol{Z}$ 的第 c 个元素可由下式计算:

$$z_c=\text{Squeeze}(\boldsymbol{X}_c)=\frac{1}{H\times W}\sum_{h=1}^{H}\sum_{w=1}^{W}\boldsymbol{X}_c(h,w)\ 。\tag{1}$$

2) 激励(Excitation)。为了完全利用上一个操作中的聚合信息,通过两层全连接层 $\boldsymbol{w}_1$、$\boldsymbol{w}_2$ 对步骤 1)得到的向量进行处理,为每个特征通道生成权重 $\boldsymbol{s}$;$\boldsymbol{w}$ 是通过学习得到的,其中 $\boldsymbol{w}_1\in\mathbf{R}^{\frac{C}{r}\times C}$,$\boldsymbol{w}_2\in\mathbf{R}^{C\times\frac{C}{r}}$,第一层

全连接层使用激活函数 ReLU(δ)进行降维得到的张量 $\boldsymbol{Z}_1(1\times1\times C/r)$,第二层全连接层使用激活函数 Sigmoid($\sigma$),进行升维得到 $\boldsymbol{X}'(1\times1\times C)$。

$$\boldsymbol{s}=\text{Excitation}(\boldsymbol{z},\boldsymbol{w})=\sigma(\boldsymbol{g}(\boldsymbol{z},\boldsymbol{w}))=\sigma(\boldsymbol{w}_2\delta(\boldsymbol{w}_1\boldsymbol{z}))。\tag{2}$$

其次,在 SFAM 中,将步骤 2)生成的权重向量 $\boldsymbol{X}'\in\mathbf{R}^{1\times1\times c}$ 作为下一步的输入张量,并将其传入到三重注意力模块的三个分支中。在第一个分支中,主要是在 H 和 C 两个维度上建立交互,首先对输入的张量 $\boldsymbol{X}_1'$ 沿 H 维度的方向逆时针旋转 90°,旋转后的张量为 $\hat{\boldsymbol{X}}_1'(1\times C\times1)$。Z-pool 的作用是将通道维度上的平均池化(AvgPool)特征和最大池化(MaxPool)特征进行汇集,并将其张量扩展到 2 维,即 $\hat{\boldsymbol{X}}_1''(1\times C\times2)$,相比 $H\times W\times C$ 可以使计算量更轻,也更能丰富表示保留的实际张量,可表示为:

$$\text{Z-pool}(\hat{\boldsymbol{X}}_1')=[\text{MaxPool}(\hat{\boldsymbol{X}}_1'),\text{AvgPool}(\hat{\boldsymbol{X}}_1')]。\tag{3}$$

经过 Z-pool 后张量 $\hat{\boldsymbol{X}}_1''$ 尺寸大小为 $1\times C\times2$,再通过一个标准卷积层(卷积核大小为 7×7、BN 层)得到 $\hat{\boldsymbol{X}}_1'''(1\times C\times1)$,通过 Sigmoid 激活函数生成一个位于 0～1 之间的归一化注意力权重值向量,最后沿 H 轴逆时针旋转 90°,得到和原始输入大小一致的张量 $\boldsymbol{X}_1$。

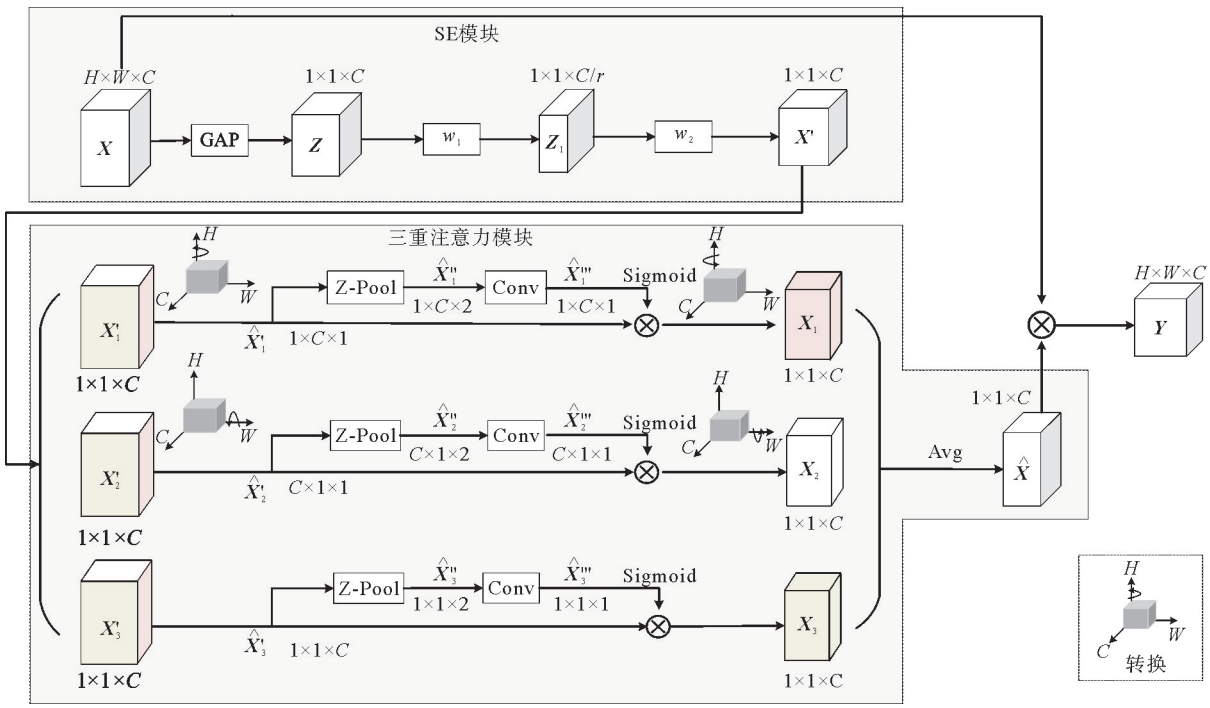

图 3　串行融合注意力(SFAM)工作原理图

Fig. 3　Operating principle diagram of serial fusion attention (SFAM)

同理,在第二个分支中,在 W 和 C 两个维度上建立交互得到张量 $\boldsymbol{X}_2$,第三个分支没有建立维度之间的交互,直接用来学习通道注意力得到张量 $\boldsymbol{X}_3$。将三个分支中得到的张量先求和再平均,得到精细张量 $\hat{\boldsymbol{X}}$ $(1\times1\times C)$,

$$\hat{\boldsymbol{X}}=\frac{1}{3}(\boldsymbol{X}_1+\boldsymbol{X}_2+\boldsymbol{X}_3)。\tag{4}$$

最后,将 $\hat{\boldsymbol{X}}$ 与原始输入张量相乘得到 SFAM 的输出矩阵 $\boldsymbol{Y}$,

$$\boldsymbol{Y}=\hat{\boldsymbol{X}}\times\boldsymbol{X}。\tag{5}$$

1.3　多尺度串行融合注意力金字塔网络

注意力金字塔是一种被广泛使用的学习多尺度特征表示的算法,可以融合不同分辨率和尺度的特征。因此,本研究在骨干网络 ResNet50-IBN-a 中加入多尺度串行融合注意力金字塔网络(SFAPNet),目的是捕

获不同尺度下图像中的显著区域,得到从“粗”到“细”的关注区域特征。

SFAPNet 的工作原理如图 4 所示。对于特征图 $\boldsymbol{Y}_{ij}$,首先进行分割操作,分割块数 n^* 依赖基数 r^*($r^*=2,3,\cdots$)的大小,即 $n^*=r^{*i}$,i 代表每层的分块数标记,每层的分割块数 n^* 会随着层数的增加呈指数增长。分割后的特征图张量 $\boldsymbol{Y}_{ij}$ 经过 SFAM,获得不同尺度、不同特征张量的判别线索,然后再将其合并,每层都得到相同大小的注意力图 A,可表示为:

$$A_i=[A_{i,j}(Y_{i,j})]_{j=1}^{n^*}。\tag{6}$$

式中,$[]_{j=1}^{n^*}$ 表示 n^* 个注意力图进行合并。这种设计会引导网络越来越关注有区分度的区域,同时前一层的输出信息会传递给下一层,从而为判别特征补充更多细粒度的特征,该过程可以通过逐元素相乘得到:

$$F_i=\alpha(A_i)*Y_{i-1}。\tag{7}$$

式中:$\alpha()$表示这些特征用归一化的注意力图进行重新加权,$*$ 表示逐元素相乘。这样从粗到细堆叠的金字塔结构,会逐步引导网络发现更多有判别性特征的重要线索。

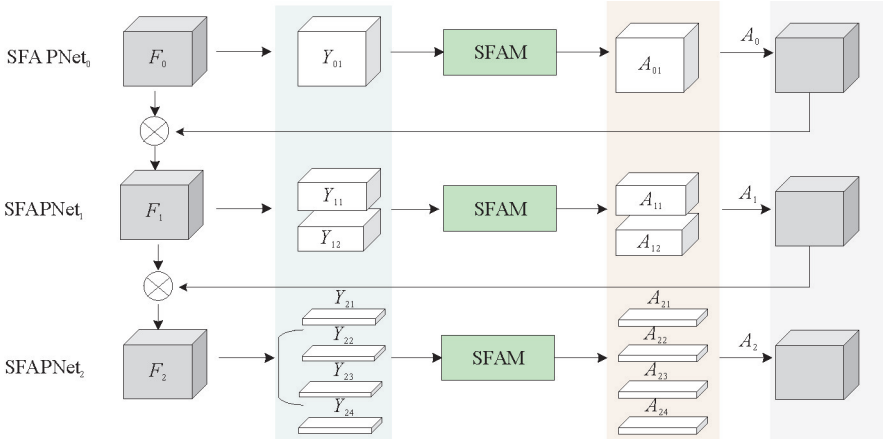

图 4 SFAPNet 工作原理图

Fig. 4 Operating principle diagram of serial fusion attention pyramid network

1.4 基于 Transformer 的新型特征校准模块

在行人重识别任务中,多尺度特征聚合可以提高网络在图像检索方面的能力。基于 HAT 结构[3],并结合基于 Transformer 的特征校准模块和深度监督聚合方法,设计了基于 Transformer 的新型特征校准模块 TFCNet,既能在深层保持语义信息,也可在浅层用细节信息丰富特征,缓解浅层特征中语义信息较少的问题,合并和保留多层次的语义信息和细节信息。

TFCNet 中的 Transformer 工作原理如图 5 所示,主要由多头自注意力层(multi-head self-attention layer,MSA)、前馈网络(feed-forward network,FFN)、归一化层(Norm)和邻域调整(neighborhood adjustment)模块组成。多头自注意力层包含 N_h 个并行的自注意力层,每一个层称为一个头(head),对于输入的特征向量 $\boldsymbol{F}_S^P$,首先在每个头中转换成查询向量 $\boldsymbol{q}\in\mathbf{R}^{(N+1)\times d}$,键向量 $\boldsymbol{k}\in\mathbf{R}^{(N+1)\times d}$ 和值向量 $\boldsymbol{v}\in\mathbf{R}^{(N+1)\times d}$,其中 $d=\dfrac{C_P}{N_h}$,N_h 为多头注意力的个数,C_P 为二维特征块的通道数,P 为特征块的大小;然后,利用 Softmax 函数对获得的注意力权重进行归一化,单头注意力及多头注意力的输出可以分别由式(8)、式(9)计算:

$$\mathrm{Attention}(\boldsymbol{q},\boldsymbol{k},\boldsymbol{v})=\mathrm{Softmax}\left(\frac{\boldsymbol{q}\boldsymbol{k}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{v},\tag{8}$$

$$\mathrm{MSA}(\boldsymbol{F}_S^P)=\mathrm{Concat}(\mathrm{head}_1,\mathrm{head}_2,\cdots,\mathrm{head}_{N_h})\boldsymbol{w}^o。\tag{9}$$

式中:d_k 为键向量$\boldsymbol{k}$的长度,除以$\sqrt{d_k}$等价于归一化操作来保证数值的稳定性;$\boldsymbol{w}^o$是训练模型所需的线

性变化矩阵,目的是使输出矩阵与输入矩阵同维。另外,每一个头的自注意力的结果会被拼接起来,使关联的特征集中在一起,更容易训练。多头自注意力层的输出和输入通过残差连接,在残差连接之后进行层归一化(LayerNorm):

$$\boldsymbol{F}_{\mathrm{S}}^{P}=\mathrm{LayerNorm}[\boldsymbol{F}_{\mathrm{S}}^{P}+\mathrm{MSA}(\boldsymbol{F}_{\mathrm{S}}^{P})]。\tag{10}$$

FFN 是一个多层感知机(multilayer perceptron,MLP)结构,由两个线性变换层和 GELU 激活函数组成,应用在 MSA 层之后:

$$\mathrm{FFN}(\boldsymbol{F}_{\mathrm{S}}^{P})=\boldsymbol{W}_{2}\gamma(\boldsymbol{W}_{1}\boldsymbol{F}_{\mathrm{S}}^{P}),\tag{11}$$

$$\boldsymbol{F}_{\mathrm{S}}^{P}=\mathrm{LayerNorm}[\boldsymbol{F}_{\mathrm{S}}^{P}+\mathrm{FFN}(\boldsymbol{F}_{\mathrm{S}}^{P})]。\tag{12}$$

式中:$\boldsymbol{W}_1$ 和 $\boldsymbol{W}_2$ 为两个线性变换层的两个参数矩阵,γ 表示 GELU 激活函数。

Transformer 的输出特征被重塑(Reshape)为与输入相同的大小。然后,这些特征被转入邻域调整模块(neighborhood adjustment module),该模块由具有批量归一化的卷积层堆叠而成。因此,TFC 的最终输出为:

$$\boldsymbol{F}_{\mathrm{S}}=\mathrm{Conv}[\mathrm{Reshape}(\boldsymbol{F}_{\mathrm{S}}^{P})]。\tag{13}$$

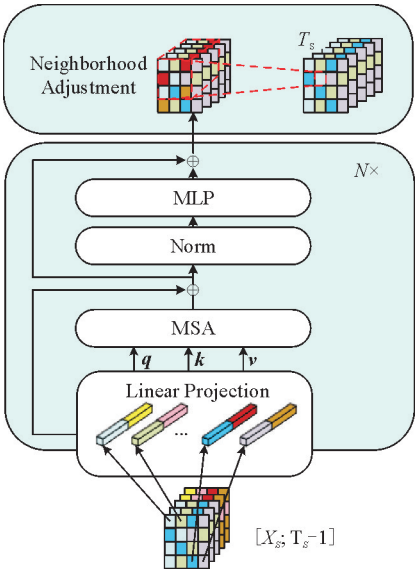

图 5　TFCNet 中的 Transformer 工作原理图

Fig. 5　Operating principle diagram of transformer in TFCNet

最后,标签平滑(label smoothing,LS)的交叉熵损失和难样本三元组损失[12]被用于端到端的训练过程。这样,通过深度监督聚合方法,Transformer 可以保留语义信息,将之前层次中挖掘的细节信息添加到当前层次中。

2　实验结果与分析

2.1　数据集、评价指标与参数设置

实验采用 3 个公开的行人重识别数据集:Market-1501[13]、DukeMTMC-reID[14]和 CUHK03[15],数据集的具体特征如表 1 所示。

通过平均精度均值(mean average precision,mAP)、平均逆负惩罚(mean inverse negative penalty,mINP)和 Rank-n 精度(第 n 次内匹配正确的概率,n=1,5,10)3 个评价指标对算法在 3 个数据集进行评估。

表 1　数据集详细信息

Table 1　Dataset details

数据集	相机个数	训练集数量	训练行人数目	测试集数量	测试行人数目	查询集数量
Market-1501	6	12 936	751	19 732	750	3 368
DukeMTMC-reID	8	16 522	702	19 889	702	2 228
CUHK03	6	7 365	767	6 732	700	1 400

采用在 ImageNet 上预训练的 ResNet50-IBN-a[9]作为实验的骨干网络,最后一个残差块的步幅设置为 1。训练时 Batch size 设定为 24,测试时 Batch size 设定为 128。训练期间使用 Adam 优化器优化网络模型。在训练过程中设置:初始学习率为 0.000 4,epoch 为 160,学习率衰减步长为[50, 70, 90, 110, 130, 150],学习率衰减因子为 0.4。除了采用学习率衰减来调整模型的学习效率外,还采用 warmup 策略,warmup 策略为线性模式,迭代次数为 100。

2.2　实验结果分析

为了测试 SFAM 的有效性,引入 ResNet50-IBN-a 作为骨干网络,并在加入 TFCNet 的基础上,将 SFAM 分别替换成 SE 和三重注意力模块,并在 Market-1501 数据集上进行评估,其中 SE 的通道缩放因子

r 为 8。从表 2 中可以看出,使用 SFAM 的性能比单独使用 SE 和三重注意力模块在 mAP、mINP、Rank-1 分别提升 0.13%、0.46%、0.27%和 0.7%、0.31%、0.15%,表明本研究提出的 SFAM 对于网络模型性能具有一定提升。

表 2 Market-1501 上三种注意力机制的实验对比

Table 2 Experimental comparison of three attention mechanisms on Market-1501 %

注意力机制	mAP	mINP	Rank-1	Rank-5	Rank-10
SE	88.97	66.89	95.40	98.52	99.32
三重注意力模块	88.40	67.04	95.52	98.55	99.29
SFAM	**89.10**	**67.35**	**95.67**	**98.72**	**99.35**

为测试 SFAM 金字塔层级数量对精度的影响,在 Market-1501 数据集上实验验证。在 ResNet50-IBN-a 的基础上添加 TFCNet,$SFAPNet_0$、$SFAPNet_1$、$SFAPNet_2$ 分别代表金字塔层级的 0、1、2 层,其中 $SFAPNet_0$ 为堆叠相同的注意力,但未分割特征图。结果如表 3 所示,实验过程中 SE 通道缩放因子 r 为 8,基数 r^* 为 2。当 SFAPNet 的层级越来越多时,精度增加会变得缓慢,这表明当金字塔的层级足够时,因为分辨率有限,无法发现更深层的判别语义信息,故选择 $SFAPNet_2$ 进行接下来的实验。

SFAPNet 每层中的分割块数是由基数 r^* 决定的,每层中分割的块数为 $r^{*i}(i=0,1,2)$。使用 $SFAPNet_2$ 作为基线模型,参数由以上实验得出的最优参数进行设置,并选择 $r^*=\{2,3,4,6\}$分别构建 SFAPNet,实验结果如表 4 所示。由表 4 可以看到,当 r^* 等于 2 时,mAP、mINP 和 Rank-1 的准确率最高,但是当 r^* 增加时,三者的准确率都会下降。

表 3 SFAPNet 层数实验结果对比

Table 3 Comparison of experimental results of SFAPNet layers %

模型	mAP	mINP	Rank-1	Rank-5	Rank-10
ResNet50-IBN-a+TFCNet	88.03	67.34	95.20	97.72	99.10
$SFAPNet_0$	88.40	67.35	95.52	98.72	99.29
$SFAPNet_1$	89.22	67.46	95.61	98.84	99.32
$SFAPNet_2$	**89.94**	**69.56**	**95.75**	**98.99**	**99.41**

为了进一步验证本研究提出的在网络中分别添加 TFCNet、SFAM 以及 $SFAPNet_2$ 的有效性,在 Market-1501 数据集进行消融实验,对比结果如表 5 所示,添加各模块之后模型的最终 mAP、mINP、Rank-1 准确率分别达到 91.00%、73.26%、96.31%,各个模块对模型性能的提升均作出重要贡献,表明本研究所提方法对于提取更加多样化的特征效果明显,具有很强的鲁棒性。

表 4 r^* 取值对比实验

Table 4 Comparison experiment of r^* values %

r^*	mAP	mINP	Rank-1	Rank-5	Rank-10
2	90.25	69.80	96.14	98.84	99.38
3	89.81	69.54	95.64	98.81	99.26
4	89.70	68.96	95.52	98.63	99.35
6	89.69	68.87	95.46	98.66	99.32

表 5 消融实验结果

Table 5 Results of ablation experiment %

方法	mAP	mINP	Rank-1	Rank-5	Rank-10
ResNet50-IBN-a	71.52	38.24	86.88	94.60	96.70
+TFCNet	88.03	67.34	95.20	97.72	99.10
+SFAM	89.10	67.35	95.67	98.72	99.35
+$SFAPNet_2$	91.00	73.26	96.31	98.97	99.47

将本研究提出的方法和近年来一些主流方法进行对比,实验结果如表 6 所示,其中“—”表示文献未给出此数据。由表 6 可知,本研究提出的方法在 3 个公开数据集上均展示了较强的鲁棒性和判别力。

3 结论

本研究提出一种基于CNN-Transformer和注意力金字塔的行人重识别网络(CTAAPNet)。引入

表 6　不同方法在三个数据集上的实验结果

Table 6　Experimental results of different methods on three datasets　　%

方法	Market-1501			DukeMTMC-reID			CUHK03		
	mAP	mINP	Rank-1	mAP	mINP	Rank-1	mAP	mINP	Rank-1
OSNet[16]	84.90	—	94.80	73.50	—	88.60	67.80/—	—	72.30/—
IGT[17]	85.28	—	94.70	76.52	—	86.49	—	—	—
SCSN[6]	88.50	—	95.70	79.00	—	90.10	80.20/83.30	—	84.10/86.30
SONA[18]	88.83	—	95.58	78.28	—	89.38	77.27/79.23	—	79.90/81.40
LAG—Net[19]	89.5	—	95.60	81.60	—	90.40	79.10/82.40	—	82.20/85.10
CTM[20]	90.20	—	96.00	82.30	—	91.60	—	—	—
本研究	**91.00**	**73.26**	**96.31**	**84.02**	**52.53**	**92.59**	**81.20/84.32**	**63.36/67.20**	**84.96/87.22**

TFC 能够更好地融合信息,并保留跨层特征中的语义信息和细节信息,通过对不同层级特征进行聚合,汇集了浅层细节信息作为深层语义信息的全局先验;设计了一种串行融合注意力机制 SFAM,结合通道和空间的信息特征,使 SFAM 和 APNet 组成 SFAPNet,对不同尺度的特征图进行由“粗”粒度到“细”粒度特征的提取,增加了模型学习差异性特征的能力;采用多重损失函数对网络模型进行监督优化,提高模型预测结果的可靠性。将提出的网络分别在 Market-1501、DukeMTMC-reID 以及 CUHK03 数据集上进行了实验验证,消融实验和对比实验结果表明,本研究提出的方法不仅可以聚合多尺度特征,还可以捕捉更精细的判别特征。本研究针对行人重识别任务面临的挑战,以深度残差网络为基准设计了行人重识别网络,尽管获得了较高的准确率,但是本研究主要用监督学习的方法开展研究,需要前期对数据集进行标注,工作量庞大。无监督学习能够直接从无标注的行人图像中学习特征,解决数据集规模小的问题,减少人工标注成本,更适应当下真实场景的行人重识别环境,下一步将会继续探索无监督行人重识别的方法。

参考文献:

[1] DOSOVITSKIY A,BEYER L,KOLESNIKOV A,et al. An image is worth 16×16 words:Transformers for image recognition at scale[C/OL]//International Conference on Learning Representations. May 3-7,2021. https://github.com/google-research/vision_transformer.

[2] PENG Z L,HUANG W,GU S,et al. Conformer:Local features coupling global representations for visual recognition[C]//Proceedingsof the IEEE/CVF International Conference on Computer Vision. 2021:367-376.

[3] 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.

[4] GUO J Y,HAN K,WU H,et al. CMT:Convolutional neural networks meet vision transformers[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022:12175-12185.

[5] 徐岩,李晓振,吴作宏,等. 基于残差注意力网络的马铃薯叶部病害识别[J]. 山东科技大学学报(自然科学版),2021,40(2):76-83.
XU Yan,LI Xiaozhen,WU Zuohong,et al. Potato leaf disease recognition via residual attention network[J]. Journal of Shandong University of Science and Technology(Natural Science),2021,40(2):76-83.

[6] CHEN X S,FU C M,ZHAO Y,et al. Salience-guided cascaded suppression network for person re-identification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020:3300-3310.

[7] LIU Z P,ZHANG L L,YANG Y. Hierarchical bi-directional feature perception network for person re-identification[C]//Proceedings of the 28th ACM International Conference on Multimedia. 2020:4289-4298.

[8] LUO H,GU Y Z,LIAO X Y,et al. Bag of tricks and a strong baseline for deep person re-identification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2019:1-10.

[9] PAN X G,LUO P,SHI J P,et al. Two at once:Enhancing learning and generalization capacities via IBN-Net[C]//Proceedings of the European Conference on Computer Vision. 2018:484-500.

[10] 李浩,樊建聪.基于多头注意力门控卷积网络的特定目标情感分析[J].山东科技大学学报(自然科学版),2022,41(2):99-107.
LI Hao,FAN Jiancong. Aspect-based sentiment analysis with multi-head attention gated convolutional network[J]. Journal of Shandong University of Science and Technology(Natural Science),2022,41(2):99-107.

[11] 王旭强,岳顺民,张亚行,等.基于注意力机制的特征融合序列标注模型[J].山东科技大学学报(自然科学版),2020,39(5):79-88.
WANG Xuqiang,YUE Shunmin,ZHANG Yaxing,et al. Attention based sequence labeling model with feature fusion[J]. Journal of Shandong University of Science and Technology (Natural Science),2020,39(5):79-88.

[12] ZHANG S Z,ZHANG Q,WEI X,et al. Person re-identification with triplet focal loss[J]. IEEE Access,2018,6(1):78092-78099.

[13] ZHENG L,SHEN L Y,TIAN L,et al. Scalable person re-identification:A benchmark[C]//Proceedings of the IEEE International Conference on Computer Vision. 2015:1116-1124.

[14] RISTANI E,SOLERA F,ZOU R,et al. Performance measures and a data set for multi-target,multi-camera tracking[C]//Computer Vision-ECCV 2016 Workshops:Amsterdam,The Netherlands,Oct. 8-10 Proceedings,Part II. Cham:Springer International Publishing,2016:17-35.

[15] LI W,ZHAO R,XIAO T,et al. Deepreid:Deep filter pairing neural network for person re-identification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014:152-159.

[16] ZHOU K Y,YANG Y X,CAVALLARO A,et al. Omni-scale feature learning for person re-identification[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019:3702-3712.

[17] 刘洋,闫冬梅,孟范伟.基于 Transformer 改进的两分支行人重识别算法[J].东北大学学报(自然科学版),2023,44(1):26-32.
LIU Yang,YAN Dongmei,MENG Fanwei,et al. Improved two-branch person re-identification algorithm based on transformer[J]. Journal of Northeastern University (Natural Science),2023,44(1):26-32.

[18] XIA B N,GONG Y,ZHANG Y,et al. Second-order non-local attention networks for person re-identification[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019:3760-3769.

[19] GONG X,YAO Z,LI X,et al. Lag-net:Multi-granularity network for person re-identification via local attention system [J]. IEEE transactions on multimedia,2021,24:217-229.

[20] 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,2022,45(6):2256-2263.
CHEN Ying,KUANG Cheng. Pedestrian re-identification based on CNN and transformer multi-scale learning[J]. Journal of Electronics & Information Technology,2022,45(6):2256-2263.

(责任编辑:傅 游)