

# 位置感知注意力及其在行人重识别中的应用

陈江萍<sup>1</sup>, 张索非<sup>2</sup>, 宋越<sup>3</sup>, 吴晓富<sup>1</sup>, 林嘉<sup>1</sup>

(1. 南京邮电大学 通信与信息工程学院, 江苏 南京 210003;

2. 南京邮电大学 物联网学院, 江苏 南京 210003;

3. 95958 部队, 上海 200120)

**摘要:**行人重识别领域的众多工作都表明,采用多分支神经网络搭配注意力模块是一种实现高性能特征嵌入的有效方式。传统方案主要关注于多分支网络结构的设计,而在注意力机制的设计上存在明显不足,如当前注意力机制缺乏对特征位置信息的有效挖掘和利用。为此,该文在多尺度特征金字塔分支 (Feature Pyramid Branch, FPB) 网络的框架下,分析了不同注意力模块的引入对系统性能的影响;在此基础上,讨论了两种在注意力机制中融入位置信息的方法,提出了一种新的位置感知注意力模块,该模块具有即插即用的优点,便于融入各种主干网络。在多个流行行人重识别标准数据集上的实验表明,融入位置感知注意力模块的 FPB 网络相比于原 FPB 网络,仅需增加 0.29 M 参数就可以显著提升最终的模型识别准确率:rank-1 在 Market1501 上提高 0.7 个百分点,在 DukeMTMC 上提高 1.5 个百分点,在 CUHK03-Labeled 上提高 2.4 个百分点,在 CUHK03-Detected 上提高 3.8 个百分点。

**关键词:**位置编码;非局部注意力模块;位置感知注意力模块;特征金字塔分支;行人重识别

中图分类号:TP391

文献标识码:A

文章编号:1673-629X(2023)01-0150-07

doi:10.3969/j.issn.1673-629X.2023.01.023

## A Novel Position-aware Attention Module and Its Use in Person Re-identification

CHEN Jiang-ping<sup>1</sup>, ZHANG Suo-fei<sup>2</sup>, SONG Yue<sup>3</sup>, WU Xiao-fu<sup>1</sup>, LIN Jia<sup>1</sup>

(1. School of Communication and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China;

2. School of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing 210003, China;

3. 95958 Troops, Shanghai 200120, China)

**Abstract:**Recent work in the field of person re-identification shows that using multi-branch neural networks is an effective way to achieve high performance feature-embedding. Traditional schemes mainly focus on the design of various efficient multi-branch network structures, but there are obvious deficiencies in the design of attention mechanism. For example, the current attention mechanism lacks the effective mining and utilization of feature position information. Therefore, we investigate the effects of using different attention modules in a multi-scale Feature Pyramid Branch (FPB) network. Then, two methods are discussed for introducing position information into attention modules, and a novel position-aware attention module is proposed, which can be used into various backbone networks. Experiments on popular person re-identification datasets show that compared with the original FPB network, the proposed FPB network with position-aware attention could achieve significantly better performance with 0.29 M parameters increase in model, the gain in the rank-1 accuracy is about 0.7% on Market1501, 1.5% on DukeMTMC, 2.4% on CUHK03-Labeled and 3.8% on CUHK03-detected.

**Key words:**position encoding; non-local block; position-aware attention module; feature pyramid branch; person re-identification

## 0 引言

行人重识别是在不同背景和视角的摄像头中匹配输入特定人的图像<sup>[1]</sup>,其在大规模人员追踪和智能视

频监控相关场景中得到了广泛应用。近年来,行人重识别取得了显著进展,但仍存在一些具有挑战性的问题,如局部遮挡、背景扰动、姿态变化等<sup>[2]</sup>。

收稿日期:2022-01-24

修回日期:2022-05-24

基金项目:国家自然科学基金资助项目(61372123, 61701252)

作者简介:陈江萍(1995-),男,硕士研究生,研究方向为深度学习;通信作者:吴晓富(1975-),男,教授,博士,研究方向为信息论与编码、计算机视觉。

最近,多尺度特征提取技术以及多分支神经网络结构已经被证明能提高行人重识别的性能,可以解决上述部分问题,如最近提出的特征金字塔分支 (Feature Pyramid Branch, FPB) 网络<sup>[3]</sup>。FPB 网络采用了一个双层特征金字塔分支,可以直接插入到骨干网中,形成一个不对称的多分支结构,同时该网络的特征金字塔分支可以从不同尺度的特征图中提取不同的特征,实验结果验证了该结构的有效性。

为了进一步提高模型的性能,各种注意力模块被用于提高主干网的特征提取能力,如位置注意力模块 (Position Attention Module, PAM)<sup>[4]</sup>。PAM 可以看作自然语言处理中多头注意力机制<sup>[5]</sup>的简化版本,利用特征间的相对位置关系输出更需要关注的行人可区分特征。文献[6]中提出的非局部注意力模块 (Non-Local Block) 通过自相关矩阵计算图片特征之间的相似性,从而输出特征之间的相关性信息。这些注意力模块可以通过特征之间的依赖关系获取行人可判别性特征,但是缺乏对特征绝对位置信息的挖掘。当行人特征间相似度较高时,仅通过特征间的相对关系无法准确识别行人。

该文在特征金字塔分支网络的框架下对比分析了

PAM 和非局部注意力模块的结构和性能,通过合理使用位置编码特征的绝对位置信息,使注意力模块可以通过特征的相对位置和绝对位置关系来区分特征,显著提升了系统的识别准确率。

主要贡献包括:

(1) 在多尺度特征金字塔分支网络框架下,分析了不同结构注意力模块实现的效率和性能;

(2) 提出了一种融合位置编码和非局部注意力机制的位置感知注意力模块,该模块具有即插即用的优点,能有效提升行人重识别网络的性能;

(3) 在多个流行行人重识别标准数据集上的实验表明,所提出的位置感知注意力模块能提升 FPB 网络的行人识别准确率,而增加的可学习参数量仅为 0.29 M。

## 1 面向行人重识别的位置感知注意力机制设计

### 1.1 特征金字塔分支 (FPB) 网络模型

采用特征金字塔分支 (FPB) 网络作为行人重识别模型的基础网络架构,该架构由全局分支和特征金字塔分支组成,如图 1 所示。

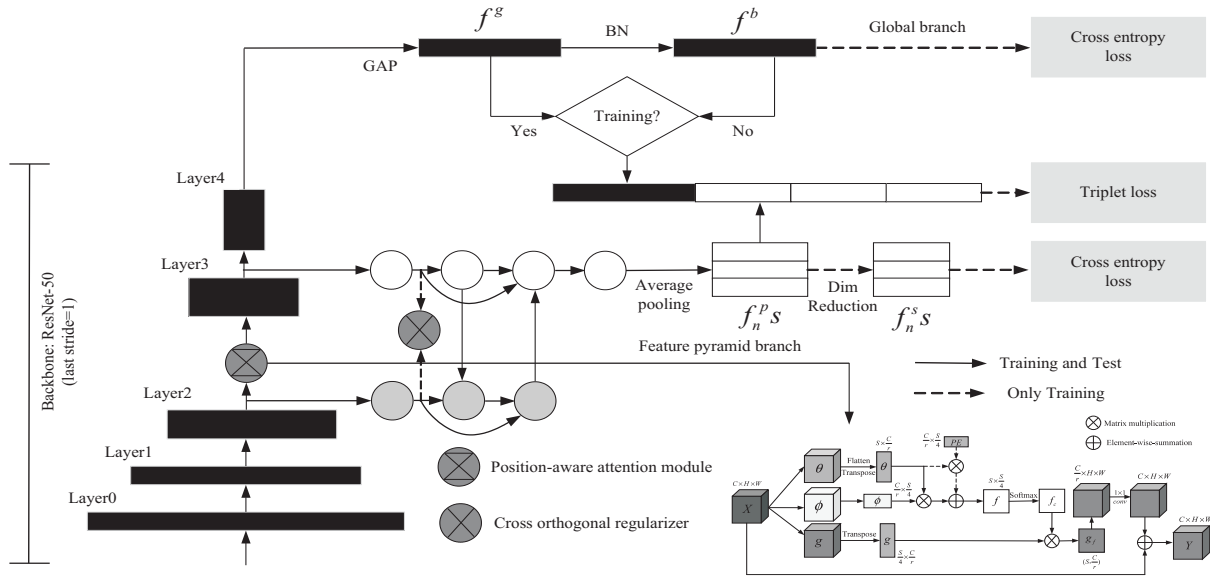


图1 融入位置感知注意力模块的金字塔特征行人重识别模型的总体架构

全局分支主要基于 Bag-Of-Tricks 方法<sup>[7]</sup>,以修正的 ResNet50 作为骨干网。与标准的 ResNet50<sup>[8]</sup>不同的是,这里去掉了 ResNet50 第四层的最后一个下采样操作,以增加输出特征图的尺寸。对主干网第四层的输出进行全局平均池化 (Global Average Pooling, GAP) 操作,再使用 BNNeck (BN), 输出一个 2 048 维向量作为全局特征向量。

除了全局分支,FPB 网络引入了一种轻量级的特征金字塔分支,用来丰富特征的多样性。这个分支的

结构是受到文献[9]的启发,将输出特征图划分为局部特征来强调局部信息。不同之处在于,特征金字塔分支是从主干网的第 2 层和第 3 层取特征 (相比文献[9],所取特征层更浅),这些浅层特征可以保留图像更多的局部细节。同时,浅层的低维特征可以减少分支中可学习参数的数量。

除此之外,FPB 网络还在骨干网第二层的输出位置插入了一个位置注意力模块,用于提取与任务有关的不同位置特征之间的相关性信息。在位置注意力模

块的基础上,通过在特征金字塔分支上加入正交正则化来进一步强化特征的多样性。如文献[3]所述,正交正则化的目的是通过降低不同通道间的特征相关性来提高特征的代表效率,对注意力模块输出特征的影响尤其明显。

在训练和测试过程中,对 FPB 网络各个分支的输出特征分别利用不同的处理策略。如图 1 所示,假设特征金字塔分支采用了平均池化后得到  $N$  个 1 024 维的特征向量,记为  $f_n^p (n \in \{1, 2, \dots, N\})$ 。经过由卷积滤波器、批处理归一化和整流线性单元激活函数组成的降维层将维数压缩到 256 维,记为  $f_n^s$ 。训练:特征金字塔分支中平均池化的输出特征  $f_n^p$  和全局分支特征  $f^p$  拼接在一起优化训练时的三元组损失;特征金字塔分支中降维层的输出特征  $f_n^s$  和经过 BN 的全局分支特征  $f^s$  分别用于优化训练时的分类损失。测试:特征金字塔分支中平均池化的输出特征  $f_n^p$  和经过 BN 的全局特征  $f^p$  拼接在一起组成了整个模型的最终特征,用于测试。

## 1.2 PAM 与非局部注意力模块的比较

注意力模块在各种深度学习场景中被证明是一种有效的机制,通过注意力模块能实现任务相关特征的有效提取。对于行人重识别任务,注意力模块的使用

有助于模型学习跟行人辨识相关的典型特征,通过让模型关注特征之间的关系,能有效提高模型的泛化能力。该文考虑以下两种注意力模块:

位置注意力模块 (Position Attention Module, PAM):该机制由文献[4]提出,看作是自然语言处理中广泛应用的多头注意力机制的简化。给定输入特征映射  $X \in R^{C \times H \times W}$ ,其中  $C$ 、 $H$  和  $W$  分别为图片通道数、高度和宽度。PAM 将图片中每个位置的特征投影和重塑到两个低维子空间和一个与输入  $X$  相同维数的空间上,得到  $Q \in R^{\frac{C}{r} \times S}$ 、 $K \in R^{\frac{C}{r} \times S}$  和  $V \in R^{C \times S}$ 。其中,  $S = H \times W$  是特征图的空间大小,  $r$  为控制子空间维数的超参数。实验表明,  $r$  取 4 时效果最好。  $X$  到  $Q$ 、 $K$  和  $V$  的投影是通过核大小为  $1 \times 1$  的 2D 卷积实现的。那么通过这个模块的输出可由公式(1)计算得出:

$$\text{attention}_p(X) = V \cdot \text{Softmax}(A_p) = V \cdot \text{Softmax}(Q^T K) \quad (1)$$

如果忽略所有的可学习参数,位置相似度矩阵  $A_p$  可以简化为一个格拉姆矩阵,可以测量  $X$  不同位置特征之间的相关性。从这个角度看,位置注意力的本质是通过每个特征与其他特征之间的相关性来重新加权特征。结构如图 2 所示,PAM 中还采用了残差结构和可学习参数  $\gamma$  来调节注意力的影响。

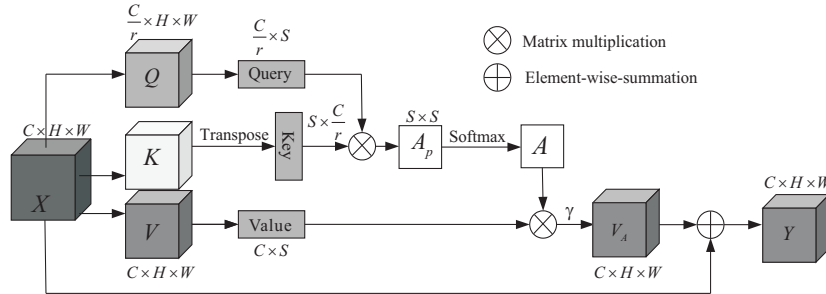


图 2 位置注意力模块

非局部注意力模块 (Non-Local Block):该模块是文献[6]中提出的自注意力模型,用于捕捉长范围的依赖关系,可以计算任意两个位置之间的关系。给定输入特征映射  $X \in R^{C \times H \times W}$ ,其中  $C$ 、 $H$  和  $W$  分别为图片通道数、高度和宽度。Non-Local Block 将图片中每个位置的特征投影和重塑到三个低维空间上,得到结果  $\theta \in R^{\frac{C}{r} \times S}$ 、 $\varphi \in R^{\frac{C}{r} \times \frac{S}{4}}$  和  $g \in R^{\frac{C}{r} \times \frac{S}{4}}$ 。其中  $S$ 、 $r$  含义和数值同 PAM。  $X$  到  $\varphi$  和  $g$  的投影是通过核大小为  $1 \times 1$  的 2D 卷积和  $2 \times 2$  的最大池化层实现的。那么通过这个模块的输出  $Y$  可由公式(2)计算得出:

$$Y_i = \frac{1}{C(X)} \sum_j f(X_i, X_j) \cdot g(X_j) \quad (2)$$

其中,  $X$  是输入信号,  $i$  是输出位置,可以是空间、时间或者时空的索引,它的响应是对所有的位置  $j$  进行枚举计算得到,  $f(X_i, X_j) = e^{\theta(X_i) \cdot \varphi(X_j)}$  用于计算  $i$  和  $j$  的相似

度,  $g(X_j) = W_g X_j$  用于计算特征图在  $j$  位置的表示,  $W_g$  是卷积核的权重参数,  $C(x) = \sum_j f(X_i, X_j)$  是归一化因子,结构如图 3 所示,删掉虚线路径时为非局部注意力模块,保留虚线路径时则为位置感知注意力模块,该结构中也采用了残差结构来调节注意力的影响。

通过将 Non-Local Block 与 PAM 进行对比,可知有如下区别:

(1) Non-Local Block 中使用  $1 \times 1$  卷积对原始图像进行特征提取时,对  $g$  分支上的特征图也进行了通道数的降维操作,并在输出前加入  $1 \times 1$  卷积进行升维到原始输入维度大小,而 PAM 中  $V$  分支没有相应的降维和升维操作;

(2) Non-Local Block 中的  $\varphi$  和  $g$  分支对应的特征图需要进行池化操作,特征图大小缩小了 4 倍,而 PAM 中  $Q$  和  $V$  分支没有池化操作;

(3) Non-Local Block 可以通过调整超参数,变种很多种形式的注意力模块,而 PAM 只是在输入特征空间(单空间)上做的 attention。

因此,可得出结论:PAM 是 Non-Local Block 的一

个特例,在特征空间上注意力机制的原理基本相同,详细的实验结果对比将在 2.3 节表 1 中详细阐述。根据实验结果,选择性能更优的 Non-Local Block 作为模型的基础注意力模块。

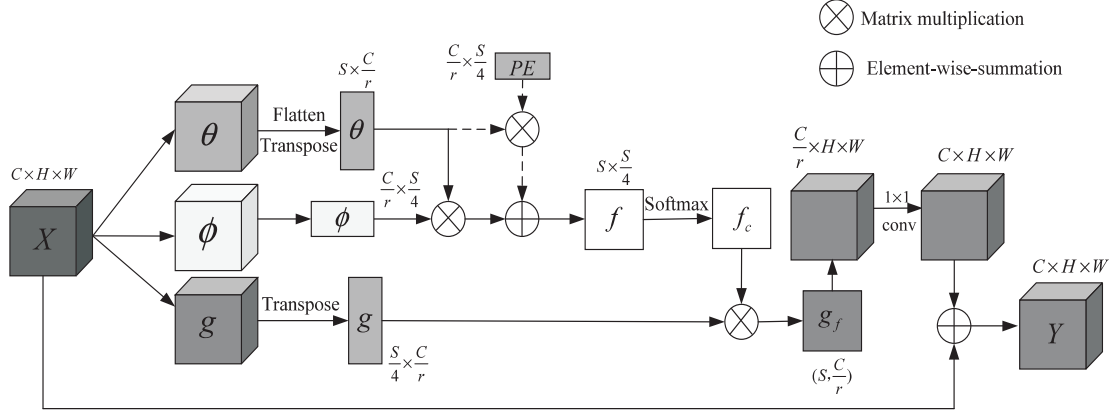


图3 注意力模块

### 1.3 位置感知注意力模块设计

为了更加充分利用特征出现的位置作为先验知识提升特征的代表性,可以在模型中加入一些特征的位置信息。

该文考虑两种方法将位置信息融入非局部注意力模块,一种是三角函数式常数位置编码 (Sinusoidal Position Encoding, SPE)<sup>[10]</sup>,另一种是可学习的位置编码 (Learned Positional Embedding, LPE)<sup>[11]</sup>。融合位置编码的非局部注意力模块—位置感知注意力模块如图 3 所示,在图中 PE 处引入位置信息,PE 指的是 LPE 或 SPE。

三角函数式常数位置编码:利用正余弦函数的周期性进行位置编码,即通过正余弦函数相乘得到相对位置关系。不同位置不同维度的编码公式如下:

$$\text{SPE}_{(p,2i)} = \sin\left(\frac{p}{10\,000^{2i/d_{\text{spe}}}}\right) \quad (3)$$

$$\text{SPE}_{(p,2i+1)} = \cos\left(\frac{p}{10\,000^{2i/d_{\text{spe}}}}\right) \quad (4)$$

其中,  $p$  代表特征位置,  $i$  代表位置向量的维度,  $d_{\text{spe}}$  是位置向量的长度。位置编码的每一个维度对应于一个正弦或余弦信号。使用三角函数设计的好处是位置  $p+k$  处单词的位置编码可以被位置  $p$  处单词的位置编码线性表示,反映两处单词的相对位置关系。

可学习的位置编码:对不同的位置随机初始化不同的位置嵌入向量,作为模型的参数进行训练,用于编码绝对位置。如图 3 中,定义一个  $\frac{C}{r} \times \frac{S}{4}$  大小的矩阵作为位置矩阵 PE,其中位置矩阵 PE 每行的初始化值从均值为 0、方差为 1 的正态分布中随机取值,随着训练过程更新。融合位置编码的非局部注意力模块内特

征与特征之间的关系可由公式(5)计算得出:

$$\mathbf{R}_{\theta,\varphi} = \theta \times \varphi \quad (5)$$

特征与位置之间的关系可由公式(6)计算得出:

$$\mathbf{R}_{\theta,\text{PE}} = \theta \times \text{PE} \quad (6)$$

其中,  $\mathbf{R}_{\theta,\varphi}$  和  $\mathbf{R}_{\theta,\text{PE}}$  均为  $S \times \frac{S}{4}$  的矩阵,将公式(5)和公式(6)相加可实现位置信息融入非局部注意力模块:

$$f = \mathbf{R}_{\theta,\varphi} + \mathbf{R}_{\theta,\text{PE}} = \theta \times \varphi + \theta \times \text{PE} \quad (7)$$

通过实验比较了两种不同方式的位置编码融入非局部注意力模块后的性能区别,以数据集 Market1501 为例,加入可学习的位置编码的 mAP 和 rank-1 分别比加入三角函数式常数位置编码高 0.2% 和 0.5%,其他数据集上的对比将在 2.4 节表 2 中详细阐述,最终选择可学习的位置编码将位置信息融入到非局部注意力模块中,形成位置感知注意力模块。同时,也可将可学习的位置编码以同样的方式应用于 PAM 中,在 Market1501 上的实验结果将在 2.4 节表 3 中详细阐述。

### 1.4 损失函数

如图 1 所示,在训练阶段,可以得到四种输出特征:  $f^{\text{e}}$ 、 $f^{\text{b}}$ 、 $f_{\text{n}}^{\text{e}}$  和  $f_{\text{n}}^{\text{s}}$ 。 $f_{\text{n}}^{\text{e}}$  和  $f_{\text{n}}^{\text{s}}$  代表平均池化后每个部分的多个特征。通过学习模型中的所有参数,优化如下的损失函数:

$$\begin{aligned} \ell_{\text{total}} = & \alpha \ell_{\text{triplet}}(f_i^{\text{e}} \odot f_{\text{n}}^{\text{e}}, y_i, f_j^{\text{e}} \odot f_{\text{n}}^{\text{e}}, y_i) + \ell_{\text{ce}}(W^{\text{b}} f_i^{\text{b}}, \\ & y_i) + \sum_{n=1}^N \ell_{\text{ce}}(W_n^{\text{s}} f_{\text{n}}^{\text{s}}, y_i) + \ell_{\text{cor}} \end{aligned} \quad (8)$$

其中,  $i$  是训练样本  $(x_i, y_i)$  的索引。 $\ell_{\text{triplet}}(\cdot)$  是在一个批次内样本  $i$  和另外一个样本  $j$  之间的困难样本三元组损失<sup>[12]</sup>。 $\odot$  表示  $f_i^{\text{e}}$  和  $f_{\text{n}}^{\text{e}}$  所有向量的拼接操作。 $\ell_{\text{ce}}(\cdot)$  是交叉熵损失<sup>[12]</sup>,  $W^{\text{b}}$  和  $W_n^{\text{s}}$  分别为  $f^{\text{b}}$  和  $f_{\text{n}}^{\text{s}}$  之后的全连接层。 $\ell_{\text{cor}}$  是  $\ell_{\text{or}}$  的交叉正交正则化版本。采用



超参数  $\alpha$  平衡不同的损失。 $\ell_{ce}(\cdot)$  的使用遵循了行人重识别模型的传统框架,而  $\ell_{triplet}(\cdot)$  通过确保不同身份样本的输出特征之间的距离大于相同身份样本的输出特征之间的距离,提高模型的泛化能力。

如公式(8)所示,损失函数包含全局分支上的特征  $f^s$  和经过 BNNeck 的特征  $f^b$ 。BNNeck 部署了一个 BN 层,通过让特征  $f^s$  经过 BNNeck,得到归一化版本  $f^b$ 。 $f^b$  在训练时用于全局分支上  $\ell_{ce}(\cdot)$  的优化,在推理时用于最终特征的一部分。而原始的  $f^s$  作为输出特征的一部分,在训练过程中对  $\ell_{triplet}(\cdot)$  进行优化。实验结果表明,仅仅在全局分支上部署 BNNeck 而不是在两个分支上部署 BNNeck 可以获得最优的性能,原因在于它会使输出的结构不对称,从而确保不同路径的特征多样性。

## 2 实验

### 2.1 数据集

该文开展了一系列的实验来分析嵌入注意力模块的性能影响,实验主要考虑了三种常用的行人重识别数据集:Market1501<sup>[13]</sup>、DukeMTMC<sup>[14]</sup>、CUHK03<sup>[15]</sup>。

Market1501 包含 32 668 张图片,由 6 个摄像头拍摄到的 1 501 个行人图片组成,每个行人都至少被两个摄像头捕获到,且在同一个摄像头中可能具有多张图片。训练集考虑了来自 751 个行人的 12 936 张图片,平均每个人有 17.2 个训练样本,测试集考虑了来自 750 个行人的 19 732 张图片。

DukeMTMC 包含 36 411 张图片,由至少两个摄像头拍摄到的 1 404 个行人图片组成,并且将仅由一个摄像头捕获到的 408 个行人的图片作为干扰物。训练集采用了 702 个行人的 16 522 张图片,测试集选用了 702 个行人的 17 661 张图片。

CUHK03 由 5 个摄像头拍摄到的 1 467 个行人的图片组成,其中 767 个行人的图片用作训练集,剩下 700 个行人的图片用作测试集。数据集包含两项:用于行人重识别的人工标注行人框图片和机器标注行人框图片。人工标注行人框的数据集有 7 368 张图片用

于训练,6 728 张图片用于测试。机器标注行人框的数据集有 7 365 张图片用于训练,7 732 张图片用于测试。

### 2.2 实现细节

骨干网络采用了在 ImageNet 上预训练过的 ResNet50,特征金字塔分支初始化选用了 He<sup>[16]</sup> 方法。训练期间采用的数据增强方法有 random horizontal flip<sup>[17]</sup>, random crop<sup>[17]</sup>, random erasing<sup>[17]</sup> 和 random patch<sup>[3]</sup>。用 Adam 优化器对模型进行了 160 轮的微调。采用线性预热策略,初始学习率设置为  $3.5 \times 10^{-5}$ ,经过 20 个 epoch 学习率达到  $3.5 \times 10^{-4}$ 。在第 60 和 90 个 epoch,学习率分别以 0.1 的速率衰减。将 Market1501、DukeMTMC 和 CUHK03 数据集的图像大小调整为  $384 \times 128$ 。实验是在 Intel E5-2680CPU 2.4 GHz 和 Nvidia Tesla P100 GPU 的硬件环境下进行的。

为了对不同方法进行定量比较,以 mAP 和 rank-1 作为标准指标。行人重识别的平均精确率 AP (Average Precision) 是指在单个行人类别上多次检索结果的平均准确率 (Precision),mAP (mean Average Precision) 是指在所有行人类别检索下的平均 AP 值;rank-1 是指搜索结果中第一张图就是正确结果的概率。所有的结果都没有使用任何重新排序或多查询融合技术。

### 2.3 PAM 与非局部注意力模块的性能对比

首先,针对 PAM 与 Non-Local Block 应用在不同的数据集上进行实验对比。在表 1 中,列出了在 Market1501、DukeMTMC、CUHK03 (Labeled) 和 CUHK03 (Detected) 上的表现。实验表明:两者在 CUHK03-Detected 上性能差别最大,mAP 和 rank-1 分别相差 0.8% 和 0.6%;在 DukeMTMC 上性能差别最小,mAP 仅相差 0.2%,rank-1 相差 0.6%;Non-Local Block 在所有数据集上的性能都优于 PAM;如果不加注意力模块,在所有的数据集上性能都将变差,在 CUHK03 表现得最明显,mAP 和 rank-1 最高分别能下降 2.7 个百分点和 2.1 百分点。

表 1 不同注意力模块的对比 %

Method	Market1501		DukeMTMC		CUHK03-Labeled		CUHK03-Detected	
	mAP	rank-1	mAP	rank-1	mAP	rank-1	mAP	rank-1
baseline+FPB	89.1	95.2	80.8	88.2	80.1	82.9	76.7	79.4
baseline+FPB+PAM	89.9	95.3	81.9	89.0	82.5	84.5	79.4	81.5
baseline+FPB+Non-Local Block	90.2	95.7	82.1	89.6	82.9	84.7	80.2	82.1

### 2.4 不同位置编码方法的对比

记 baseline+FPB+Non-Local Block 为 FPB\*。将三角函数式常数位置编码和可学习的位置编码分别加

入表 1 中的 Non-Local Block,在表 2 中列出了这两种方法的性能比较。实验表明:两者在 CUHK03-Labeled 上性能差别最大,mAP 和 rank-1 分别相差

1%和0.4%;在Market1501上性能差别最小,mAP和rank-1分别仅相差0.2%和0.5%;可学习的位置编码在所有数据集上的性能都优于三角函数式常数位置编码;如果不加可学习的位置编码,在所有数据集上的性

能都会有所下降,在CUHK03-Labeled上表现得最明显,mAP最高下降0.9百分点,rank-1下降0.6百分点。因此认为:在行人重识别中,编码特征的绝对位置比编码特征的相对位置更有效。

表2 不同位置编码方法的对比 %

Method	Market1501		DukeMTMC		CUHK03-Labeled		CUHK03-Detected	
	mAP	rank-1	mAP	rank-1	mAP	rank-1	mAP	rank-1
FPB *	90.2	95.7	82.1	89.6	82.9	84.7	80.2	82.1
FPB * +SPE	90.0	95.4	82.0	89.0	82.8	84.9	79.9	82.6
FPB * +LPE	90.2	95.9	82.3	89.7	83.8	85.3	80.4	83.2

## 2.5 注意力机制和位置编码方法的消融对比

以FPB网络为框架,嵌入位置感知注意力模块,构造了最终的模型。表3列出了在Market1501数据集上每次尝试的影响。首先,列出了骨干网ResNet50结合文献[7]中技巧的性能,记为baseline,证明了特征金字塔分支(FPB)的有效性。然后,分别加入不同的注意力模块PAM和Non-Local Block进行对比。最后,将可学习的位置编码分别融合到PAM和Non-

Local Block中。

从表3中可以看到,加入注意力模块PAM或者Non-Local Block性能都得到了提升,但加入Non-Local Block性能更好,mAP和rank-1比加入PAM高0.3%和0.4%;最后分别在PAM和Non-Local Block中加上可学习的位置编码,性能进一步提升,但性能最好的还是在Non-Local Block中加入可学习位置编码。

表3 注意力机制和位置编码方法的消融对比

Method	Params/M	Market1501	
		mAP/%	rank-1/%
baseline	25.05 M	84.6	94.0
baseline+FPB	29.04 M	89.1	95.2
baseline+FPB+PAM	29.44 M	89.9	95.3
baseline+FPB+PAM+LPE	29.54 M	90.0	95.5
baseline+FPB+Non-Local Block	29.31 M	90.2	95.7
baseline+FPB+Non-Local Block+LPE	29.33 M	90.2	95.9

## 2.6 与其他算法的横向比较

下面给出所提出的融入位置感知注意力模块的FPB网络和其他算法在不同数据库的性能比较。在表4中,列出了在Market1501, DukeMTMC, CUHK03

(Labeled)和CUHK03(Detected)这四个任务中的表现。从表4中可以看到,将位置感知注意力模块嵌入到FPB网络后,性能得到显著提高,与表中的其他算法相比,在上述几个数据集上性能都是较高的。

表4 与其他算法的横向比较 %

Method	Market1501		DukeMTMC		CUHK03-Labeled		CUHK03-Detected	
	mAP	rank-1	mAP	rank-1	mAP	rank-1	mAP	rank-1
PCB+RPP <sup>[9]</sup>	81.6	93.8	69.2	83.3	-	-	57.5	63.7
Bag-Of-Tricks <sup>[7]</sup>	85.9	94.5	76.4	86.4	-	-	-	-
ABD-Net <sup>[18]</sup>	88.28	95.6	78.59	89.0	-	-	-	-
Pyramid <sup>[19]</sup>	88.2	95.7	79.0	89.0	76.9	78.9	74.8	78.9
Adaptive L2 <sup>[20]</sup>	88.9	95.6	81.0	90.2	-	-	-	-
CDNet <sup>[21]</sup>	86.0	95.1	76.8	88.6	-	-	-	-
L3DS <sup>[22]</sup>	87.3	95.0	76.1	88.2	-	-	-	-
PAT <sup>[23]</sup>	88.0	95.4	78.2	88.8	-	-	-	-
HOReID <sup>[24]</sup>	90.00	96.45	81.03	89.12	-	-	-	-
FPB <sup>[3]</sup>	89.1	95.2	80.8	88.2	80.1	82.9	76.7	79.4
FPB *	90.2	95.7	82.1	89.6	82.9	84.7	80.2	82.1
FPB * +LPE	90.2	95.9	82.3	89.7	83.8	85.3	80.4	83.2

### 3 结束语

该文提出了一种新的位置感知注意力模块,该模块使用可学习的位置编码学习特征图中子特征间的位置关系,提升了特征的代表性。同时该模块具有即插即用的优点,能方便融入到行人重识别主干网。实验结果表明:在FPB网络中嵌入位置感知注意力模块,增加的计算量小,但能显著提高模型的识别能力。

#### 参考文献:

- [1] 刘颖,武阳阳,李娜. 基于深度学习的行人属性识别综述[J]. 西安邮电大学学报,2021,26(2):62-69.
- [2] 李擎,胡伟阳,李江昀,等. 基于深度学习的行人重识别方法综述[J/OL]. 工程科学学报,2021:1-13.
- [3] ZHANG S, YIN Z, WU X, et al. FPB: feature pyramid branch for person re-identification[J]. arXiv:2108.01901, 2021.
- [4] 廖开翔. 基于位置注意力机制的多模态学习方法研究[D]. 西安:西安电子科技大学,2020.
- [5] LIU Z, LIN Y, CAO Y, et al. Swin transformer: hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF international conference on computer vision. Los Alamitos: IEEE, 2021: 10012-10022.
- [6] WANG X, GIRSHICK R, GUPTA A, et al. Non-local neural networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Salt Lake City: IEEE, 2018: 7794-7803.
- [7] LUO H, GU Y, LIAO X, 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. Long Beach: IEEE, 2019.
- [8] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas: IEEE, 2016: 770-778.
- [9] 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 (ECCV). Germany: Springer, 2018: 480-496.
- [10] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Advances in neural information processing systems. California: NeurIPS, 2017: 5998-6008.
- [11] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: transformers for image recognition at scale[J]. arXiv:2010.11929, 2020.
- [12] HERMANS A, BEYER L, LEIBE B. In defense of the triplet loss for person re-identification[J]. arXiv: 1703.07737, 2017.
- [13] ZHENG L, SHEN L, TIAN L, et al. Scalable person re-identification: a benchmark[C]//Proceedings of the IEEE international conference on computer vision. Washington, DC: IEEE, 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]//European conference on computer vision. Germany: Springer, 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. Columbus: IEEE, 2014: 152-159.
- [16] HE K, ZHANG X, REN S, et al. Delving deep into rectifiers: surpassing human-level performance on imagenet classification[C]//Proceedings of the IEEE international conference on computer vision. Washington, DC: IEEE, 2015: 1026-1034.
- [17] ZHONG Z, ZHENG L, KANG G, et al. Random erasing data augmentation[C]//Proceedings of the AAAI conference on artificial intelligence. Palo Alto, CA: AAAI, 2020: 13001-13008.
- [18] CHEN T, DING S, XIE J, et al. Abd-net: attentive but diverse person re-identification[C]//Proceedings of the IEEE/CVF international conference on computer vision. Seoul: IEEE, 2019: 8351-8361.
- [19] ZHENG F, DENG C, SUN X, et al. Pyramidal person re-identification via multi-loss dynamic training[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Seoul: IEEE, 2019: 8514-8522.
- [20] NI X, FANG L, HUTTUNEN H. Adaptive l2 regularization in person re-identification[C]//2020 25th international conference on pattern recognition (ICPR). [s. l.]: IAPR, 2021: 9601-9607.
- [21] LI H, WU G, ZHENG W S. Combined depth space based architecture search for person re-identification[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. [s. l.]: IEEE, 2021: 6729-6738.
- [22] CHEN J, JIANG X, WANG F, et al. Learning 3D shape feature for texture-insensitive person re-identification[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. [s. l.]: IEEE, 2021: 8146-8155.
- [23] LI Y, HE J, ZHANG T, et al. Diverse part discovery: occluded person re-identification with part-aware transformer[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. [s. l.]: IEEE, 2021: 2898-2907.
- [24] WANG P, ZHAO Z, SU F, et al. HOREID: deep high-order mapping enhances pose alignment for person re-identification[J]. IEEE Transactions on Image Processing, 2021, 30: 2908-2922.