doi：10.3969/j.issn.2095-0411.2023.02.008

基于金字塔分割和时空注意力的视频行人重识别

王洪元，徐志晨，陈海琴，丁宗元，李鹏辉

（常州大学 计算机与人工智能学院，江苏 常州 213164）

摘要：针对视频行人重识别任务中存在的行人外观、遮挡等问题，研究并设计了一个基于金字塔分割和注意力机制的视频行人重识别模型。首先，为了增强图模型对行人局部特征的识别能力，提出了多个尺度的水平金字塔分割方法，将各特征分别分割成不同大小的区域，并池化成统一尺寸后输入图模型。另外，鉴于简单的时空注意模块容易因遮挡破坏行人特征，因此使用时空相关注意力方法改进时空注意模块，逐步学习并聚合空间局部信息，同时在时序上相互作用，抑制行人干扰特征并增强判别特征。将模型在 Mars 和 DukeMTMC-VideoReID 两个数据集上进行了评估，实验结果证实了文中提出方法的有效性。

关键词：视频行人重识别；深度学习；图模型；注意力机制；加权损失策略

中图分类号：TP 391.4　　**文献标志码**：A　　**文章编号**：2095-0411(2023)02-0066-11

Video-based person re-identification based on pyramid segmentation and spatial-temporal attention

WANG Hongyuan, XU Zhichen, CHEN Haiqin, DING Zongyuan, LI Penghui

(School of Computer Science and Artificial Intelligence, Changzhou University, Changzhou 213164, China)

Abstract: Aiming at the problems of similar appearance and occlusion of people in the video person re-identification, a video-based person re-identification model based on pyramid segmentation and attention mechanism was studied and designed. First, in order to enhance the recognition ability of the graph model for the local features of pedestrians, a multi-scale horizontal pyramid segmentation method was proposed. In addition, given that the simple spatiotemporal attention module was prone to damage person features due to occlusion, the spatiotemporal attention module was improved using the spatiotemporal correlation attention method, which gradually learns and aggregates spatially local information while interacting in time sequence to suppress person interference features and enhance discriminative features. This paper evaluates the model on Mars and DukeMTMC-VideoReID datasets, and the experimental results confirm the effectiveness of the proposed method.

收稿日期：2022-10-29。

基金项目：国家自然科学基金资助项目(61976028，61572085，61070121)。

作者简介：王洪元(1960—)，男，江苏常熟人，博士，教授。E-mail：hywang@cczu.edu.cn

引用本文：王洪元，徐志晨，陈海琴，等. 基于金字塔分割和时空注意力的视频行人重识别[J]. 常州大学学报(自然科学版)，2023，35(2)：66-76.

Key words：video-based person re-identification；deep learning；graph model；attention mechanism；weighted loss strategy

随着社会的不断进步和经济的飞速发展，城市的公共安全问题也受到了越来越多的关注。近年来，为了建立安防智能化现代城市，摄像头被广泛安装在街道、校园、商城等人员密集的公共场所[1]。摄像头构建了一个可靠的保护系统[2]，视频监控[3]也成为维护治安的重要技术手段。目前，视频监控技术主要依靠人力分析，并结合有限的智能化方法[4]减轻人们的工作量。然而，城市道路交错繁多，面对摄像监控每时每秒获取到的视频数据[5]，依靠目前的办法分析数据往往需要耗费大量的人力资源，耗时耗力且效率低下。因此，如何能够高效率分析视频数据，搭建智能监控系统[6]，是公共安全领域内的重要任务[7]。

一个完整的智能监控系统应当能够实现 3 个功能：对指定监控目标的自动检测，后续的跟踪及跨摄像头数据下的再识别[8]。行人重识别作为该系统中关键的一环，目的在于能够快速匹配不同摄像头拍摄的不同场景下的目标行人，定位该目标的行程轨迹[9]。具体而言，如图 1 所示，摄像头 1 和摄像头 2 的监控区域视野不重合，从摄像头 1 采集数据中捕获目标行人作为需要查询的对象，从摄像头 2 采集到的行人组成候选人员名单，将待查询对象与候选对象逐个匹配，通过衣着、外观、体型、姿态等特征，识别出与查询对象相似度最高的行人，从而实现跨摄像头的不同场景下的目标行人的匹配与跟踪[10]。课题组在之前的工作中[11]，提出了一种基于图神经网络的方法，一方面，构建了特征关系图，挖掘了不同帧内不同节点之间的关联信息，为图模型提供了时空信息。另一方面，使用分块结构和全局结构两个分支挖掘互补信息。最后，为了弥补度量学习中样本信息丢失过多问题，采用加权对比损失策略，为每个样本分配一个连续分布的分数，充分利用小批次中的样本信息。在后续的研究中，作者发现了之前工作中的不足，首先，注意到图分支中，在挖掘局部特征的过程中，可以同时引入全局特征，增强模型从整体到局部挖掘信息的性能。其次，之前使用的时空注意模块仅通过简单的转置、卷积操作生成注意图，容易丢失大量时间信息，且没有结合学习图像间的时间信息与图像内区域信息。另外，发现在训练阶段使用的策略有所不足。

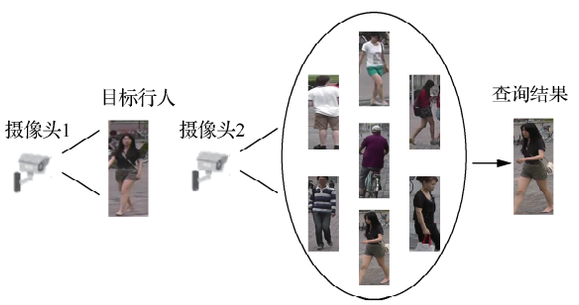


图 1　行人重识别任务应用示例

Fig.1　Application example of person re-identification task

基于以上研究，文章提出了水平金字塔分割的方法，将图片特征分别水平分割成不同区域，有效增强图模型对行人整体到局部特征的识别能力。另外，使用时空相关注意力方法改进时空注意模块，联合局部和全局信息及依赖性，挖掘并互补图片时间和空间相关性信息，学习视频行人的时空特征。最后，在 Mars 和 DukeMTMC-VideoReID 两个数据集上，将先前的方法在本文的实验设置参数下重新训练模型作为 Baseline，并统一使用交叉熵损失和三元组损失作为目标损失函数。

总的来说，文章①提出了多个尺度的水平金字塔分割的方法，将图片特征水平分割成多个区域，有效增强图模型对行人整体到局部特征的识别能力；②使用时空相关注意力方法作为时空注意模块，联合局部和全局信息及依赖性，挖掘并互补图片时间和空间相关性信息，学习视频行人的时空特征。在 Mars 和 DukeMTMC-VideoReID 数据集上，使用同一个框架和训练策略，比较了先前工作模型和本文提出的改进模型，结果证明了本文方法的有效性。

1 相关工作

1.1 图神经网络

现有的视频识别方法主要侧重于挖掘视频中丰富的时空线索。对时空线索，大多数研究采用平均池化或加权策略来融合框架特征。对时间线索，现有的方法使用光流、递归神经网络、3D 卷积或者非局部块来建模时间关系。最近，文献［12］提出联合捕捉短期和长期的时间关系。

目前，图神经网络及其变体优秀的关系建模能力，已成功应用于人体动作识别，视频分类和多标签图像识别等计算机视觉任务中。在行人重识别领域中引入了图网络模型相关的方法，也进一步提升了行人重识别的性能。在文献［13］中，作者将图注意力网络（Graph Attention Network，GAT）与特征提取网络结合在一起，从时空域的视频序列中提取具有判别性的特征并使网络专注于这些优秀的特征区域，再通过时空图发现帧与区域间的关系变化，学习特征图中的权重矩阵。同样，WU 等[14]介绍了一种图神经网络，通过利用姿态对齐和特征亲和力关系 2 个分支实现相关区域特征之间的关联，构建有判别性和鲁棒性的视频特征表示。YANG 等[15]提出了一种新颖的时空图卷积网络，构建空间和时间 2 个分支。空间分支提取人体形成的结构，时间分支从相邻帧中挖掘判别线索，联合 2 个分支提取与外观信息互补的时空信息，有效克服视觉相似的负样本的遮挡问题和视觉模糊问题。LIU 等[16]利用卷积神经网络和人体关键点估计从多尺度提取行人语义局部特征，并设计了新颖的三维图卷积和跨尺度图卷积，解决跨时空和跨尺度信息的传播问题，获取视频行人的结构信息和时空特征。

1.2 注意力机制

注意力机制的方法大部分都是将时间注意和空间注意分开学习，忽视了两者之间的关系。对此，CHEN 等[17]提出了一种联合注意力时空特征聚合网络（Joint Attentive spatial-temporal Feature aggregation Network，JAFN），通过质量感知和帧感知 2 个注意力模块，衡量图像区域的质量以及图像帧在时序中的重要性。ZHU 等[18]也针对同样的问题，提出了一个自适应的时空注意力网络，首先挖掘出每个输入帧的空间语义关系以及帧之间的时间依赖关系，再利用多个自适应时空融合模块在多级特征图上获取更精确的时空注意力。除了设计时间和空间维度上的注意力模块，自注意力机制也被应用在行人重识别任务中。ZHANG 等[19]设计了一个自协作注意力网络，利用每一帧特征与视频特征的相关性，增强更为关键的帧，获取更好的时序特征表示。WANG 等[20]利用时空参考注意挖掘时间和空间特征，并用金字塔时空聚合框架逐步聚合得到最终的视频表示。HOU 等[21]提出了一个新颖的 BiCnet 网络，通过多个平行的注意力模块发现并互补不同帧中的行人空间特征。另外同时自适应地捕捉帧长时间与短时间的动态关系，长时间帧提供更多不同的信息用来解决遮挡问题，短时间帧获取行人动作姿态用来解决快速移动问题。

2 方 法

针对先前提出方法中存在的问题，进行了持续研究与并提出了改进方法。首先，为了增强图模型对行人局部特征的识别能力，使用了 3 个尺度的水平金字塔分割的方法，将图片特征分割成 1 块，2 块和 4 块，总共 7 块区域，并用平均池化的方法统一每块区域的尺寸大小，便于输入后续的图神经网络。其次，使用了时空相关注意力逐步学习并聚合空间局部信息，同时在时序上相互作用，获得全局依赖性，抑制行人干扰性特征并增强有判别性特征。将先前工作的模型架构进行改进，结构如图 2 所

示。另外，为了达到更好的模型性能，更新了训练时使用的策略及参数。同时，为了与先前工作的方法进行比较，首先将先前工作的方法利用更新的参数重新进行了模型训练作为Baseline，并统一使用交叉熵损失和三元组损失作为目标损失函数。文章将在下面的段落逐个介绍相关模块。

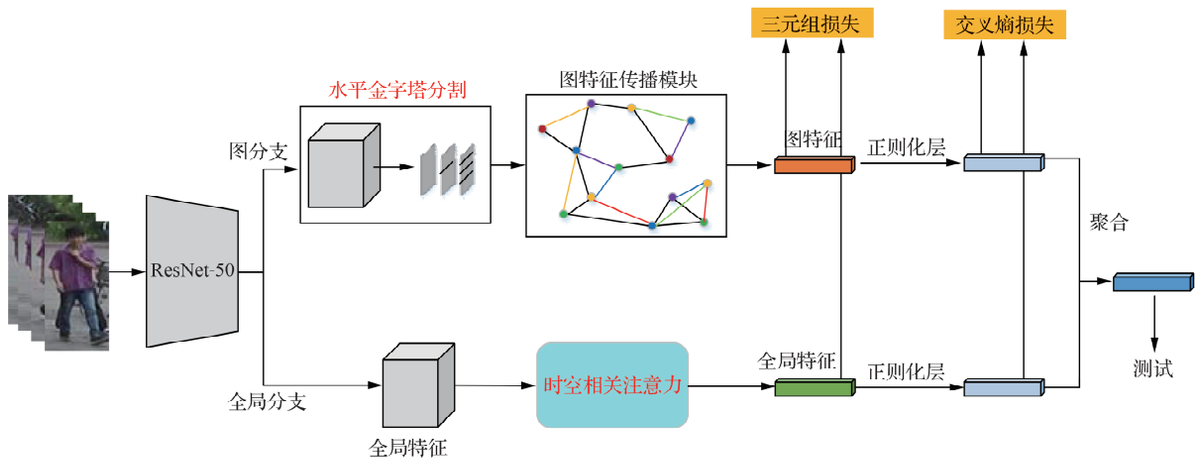


图 2　本文方法的整体框架

Fig.2　Overall architecture of our proposed method

2.1　水平金字塔分割

由于卷积神经网络的全连接层输入维度大小是固定的，为了应对这样的限制条件，HE 等[22]提出了空间金字塔池化网络，无论输入维度大小，都能够生成固定维度的向量作为输出，并在局部空间通过池化的方式保持区域位置的信息不被破坏。同时，多级空间池化方法也被证明能够增强变形后对象的鲁棒性，提高目标分类和检测任务的模型性能。同样的，文献［23］中提出的金字塔池化模块也有类似的效果，金字塔层级池化将特征图分成多级不同的子区域，并通过池化的方式得到不同的特征表示。

根据上述方法，文章设计了水平金字塔分割的方法，为了增强行人在各个尺度上的局部特征，学习有判别性的特征。由于人的布局是上下分布，从头到脚，因此该方法将特征图以水平分割的方式分成多个区域，结构如图 3 所示。具体而言，特征提取网络输出得到特征图 $\{\boldsymbol{F}_t\}_{t=1,\cdots,T}$（正整数 T 对应输入视频的帧数），特征图 $\boldsymbol{F}_t$ 的大小为高（H）×宽（W）。在水平金字塔分割中采用 M（M 为正整数）个金字塔尺度，根据不同的尺度将特征图 $\boldsymbol{F}_t$ 水平平均分割成 2^{M-1} 个空间 $\boldsymbol{F}_{i,j}^t$，其中，$i=1,\cdots,M$ 表示金字塔尺度，$j=1,\cdots,2^{M-1}$ 表示在该金字塔尺度下分成的第 j 个区域。例如，$\boldsymbol{F}_{3,4}^t$ 表示第 3 个金字塔尺度下的第 4 块空间，每个空间的高和宽乘积大小为 $H\times W/2^{M-1}$。然后，通过一个全局平均池化和一个卷积层，将每个分割后的空间 $\boldsymbol{F}_{i,j}^t$ 统一至同一大小维度，这样，共获得 $N=2^0+2^1+\cdots+2^{M-1}$ 个向量 $\boldsymbol{G}_{i,j}^t$

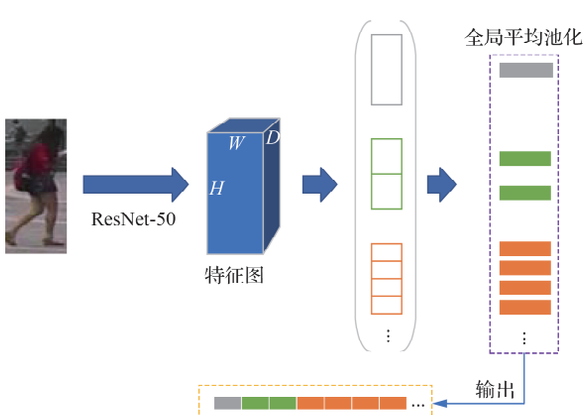


图 3　水平金字塔分割结构

Fig.3　Horizontal pyramid segmentation structure

$$\boldsymbol{G}_{i,j}^t=\mathrm{Conv}(\mathrm{avgpool}(\boldsymbol{F}_{i,j}^t)) \tag{1}$$

式中：avgpool 为平均池化；Conv 为一个卷积层，用来连接输出特征。通过这种方式，可以从全局到局部，从粗到细捕捉人物部位的判别能力。

2.2 时空相关注意力

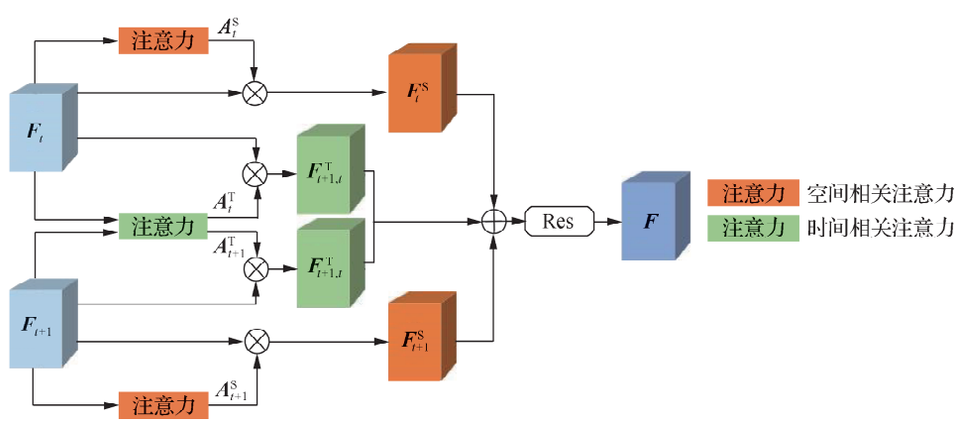
目前，注意力机制已经被广泛应用于行人重识别任务，通过注意力关注更多行人信息，从而获取有判别性的特征表示。在先前工作中也使用到了时空注意模块，模型也有明显的提升，但存在没有同时学习图像帧空间内和时序上的特征，缺少时间上长期和短期关系的相互作用的问题，限制了模型挖掘视频潜在信息的能力，这也是目前大多数基于注意力模块的问题。对此，文章使用了一个时空相关注意力模块，通过图像空间信息增强表征信息，再利用时序上的关系增强行人的特征信息，同时能够抑制干扰信息增强行人的判别特征。时空相关注意力模块架构如图 4（a）所示，该架构主要由时间相关注意和空间相关注意 2 个模块组成，时间相关注意和空间相关注意的架构图分别如图 4（b）和图 4（c）所示。在图片序列 $\{\boldsymbol{I}_t\}_{t=1,\cdots,T}$ 通过特征提取器得到特征图 $\{\boldsymbol{F}_t\}_{t=1,\cdots,T}$ 后，任意选取一对时间相邻的特征图 $\{\boldsymbol{F}_t, \boldsymbol{F}_{t+1}\}$ 作为时空相关注意力模块输入，通过时间相关注意模块和空间相关注意模块可以得到 $\boldsymbol{A}_t^{\mathrm{S}}$，$\boldsymbol{A}_{t+1}^{\mathrm{S}}$，$\boldsymbol{A}_t^{\mathrm{T}}$，$\boldsymbol{A}_{t+1}^{\mathrm{T}}$ 4 个注意力图后，利用矩阵计算得到细化的特征图，计算式为

$$\boldsymbol{F}_t^{\mathrm{S}}=\boldsymbol{A}_t^{\mathrm{S}} \circ \boldsymbol{F}_t,\ \boldsymbol{F}_{t,t+1}^{\mathrm{T}}=\boldsymbol{A}_t^{\mathrm{T}} \circ \boldsymbol{F}_t \tag{2}$$

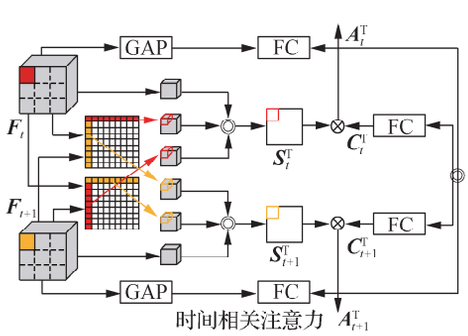
式中 ∘ 为哈达玛积。之后，将细化的特征图元素相加，并通过残差块得到融合特征。最终特征为

$$\boldsymbol{F}=\mathrm{Res}\left[\left(\boldsymbol{F}_{t,t+1}^{\mathrm{T}}+\boldsymbol{F}_{t+1,t}^{\mathrm{T}}\right)+\left(\boldsymbol{F}_t^{\mathrm{S}}+\boldsymbol{F}_{t+1}^{\mathrm{S}}\right)\right] \tag{3}$$

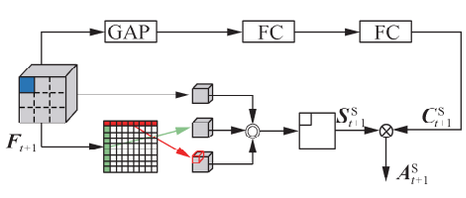
式中 Res 为文献［24］中提出的一个残差块。



（a）时空相关注意力架构图



（b）时间注意力架构图



（c）空间注意力架构图

说明：GAP—全局平均池化；FC—全连接层；Res—残差块；◎—级联。

图 4 时空相关注意力整体与局部架构图

Fig.4 Space-time related attention overall and local architecture diagram

2.2.1　空间相关注意力

目前，特征提取网络 Resnet-50 已经能够很好地提取图片特征，但由于在特征融合的过程中，背景等噪声的影响容易对行人特征造成影响，因此，行人的空间信息在行人重识别任务中起着关键作用。为此，使用空间相关注意力增强行人所在区域信息并抑制干扰信息。受文献［25］启发，将注意力图 $\boldsymbol{A}_t^{\mathrm{S}}$ 拆分为 2 个低维度注意力掩码，分别代表空间注意掩码 $\boldsymbol{S}_t^{\mathrm{S}}$ 和通道注意掩码 $\boldsymbol{C}_t^{\mathrm{S}}$

$$\boldsymbol{A}_t^{\mathrm{S}}=\boldsymbol{S}_t^{\mathrm{S}}\circ\boldsymbol{C}_t^{\mathrm{S}} \tag{4}$$

式中：$\boldsymbol{S}_t^{\mathrm{S}}\in R^{1\times H\times W}$；$\boldsymbol{C}_t^{\mathrm{S}}\in R^{C\times 1\times 1}$。同时在空间相关注意力中引入了 2 个分支，分别学习空间注意掩码和通道注意掩码，结构如图 4（c）。

在空间注意掩码学习过程中，按照文献［26］的方法，将输入的特征图 $\boldsymbol{F}_i$ 看作具有 $N=H\times W$ 个节点的二维图，再给每个节点重新分配编号为 1，…，N。这样特征图就可以表示为 N 个特征节点 $\boldsymbol{x}_i\in R^C$，$i=1$，…，N。之后，节点 i 和节点 j 的关系可以定义为

$$\boldsymbol{r}_{i,j}^{\mathrm{S}}=\theta^{\mathrm{S}}(\boldsymbol{x}_i)^{\mathrm{T}}\varphi^{\mathrm{S}}(\boldsymbol{x}_j) \tag{5}$$

式中：θ^{S}，φ^{S} 是 2 个由 1×1 空间卷积层、BN 正则化层和 ReLU 激活层组成的嵌入函数，$\theta^{\mathrm{S}}(\boldsymbol{x}_i)=\mathrm{ReLU}(W_\theta\boldsymbol{x}_i)$，$\varphi^{\mathrm{S}}(\boldsymbol{x}_i)=\mathrm{ReLU}(W_\varphi\boldsymbol{x}_i)$，其中，$W_\theta$ 和 $W_\varphi\in R^{\mathrm{S}/s_1\times C}$，$s_1$ 是一个预设的正整数，用来控制降维比例，$\theta^{\mathrm{S}}(\boldsymbol{x}_i)^{\mathrm{T}}$ 中的 T 表示转置运算。类似地，节点 j 和节点 i 的关系可以表示为 $\boldsymbol{r}_{j,i}^{\mathrm{S}}$。用该方法建立图内所有节点的关系，对于节点 i，可以定义关系向量为

$$\boldsymbol{r}_i^{\mathrm{S}}=\gamma^{\mathrm{S}}(\boldsymbol{r}_{i,1}^{\mathrm{S}},\cdots,\boldsymbol{r}_{i,N}^{\mathrm{S}},\boldsymbol{r}_{1,i}^{\mathrm{S}},\cdots,\boldsymbol{r}_{N,i}^{\mathrm{S}}) \tag{6}$$

式中 γ^{S} 同样是一个嵌入函数，与 θ^{S} 结构相同。如式（7）所示，通过堆叠关系向量 $\boldsymbol{r}_i^{\mathrm{S}}$ 和节点 $\boldsymbol{x}_i$ 的嵌入向量 $\boldsymbol{\beta}(\boldsymbol{x}_i)$，能够得到节点 $\boldsymbol{x}_i$ 所在位置的关系值向量，$\boldsymbol{\beta}(\boldsymbol{x}_i)$ 与 $\boldsymbol{r}_{j,i}^{\mathrm{S}}$ 的向量格式一致

$$\boldsymbol{v}_i^{\mathrm{S}}=[\boldsymbol{r}_i^{\mathrm{S}},\boldsymbol{\beta}(\boldsymbol{x}_i)] \tag{7}$$

最后，得到所有节点的关系值向量，更新每个节点位置特征，构建关系值矩阵 $\boldsymbol{V}^{\mathrm{S}}$，并通过卷积层和 Sigmoid 激活函数得到空间注意掩码

$$\boldsymbol{S}^{\mathrm{S}}=\mathrm{Sigmoid}(\mathrm{Conv}(\boldsymbol{V}^{\mathrm{S}})) \tag{8}$$

在通道注意掩码学习过程中，首先通过平均池化处理输入的特征图 $\boldsymbol{F}_i$

$$\boldsymbol{X}_i^{\mathrm{S}}=\frac{1}{H\times W}\sum_{h=1}^{H}\sum_{w=1}^{W}\boldsymbol{f}_{w,h} \tag{9}$$

再通过 2 个连续的全连接层得到通道注意掩码 $\boldsymbol{C}^{\mathrm{S}}\in R^{C\times 1\times 1}$

$$\boldsymbol{C}^{\mathrm{S}}=\mathrm{Sigmoid}(\mathrm{F}_{\mathrm{C2}}(\mathrm{F}_{\mathrm{C1}}(\boldsymbol{X}_i^{\mathrm{S}}))) \tag{10}$$

式中 F_{C1}，F_{C2} 分别为全连接层 1 和全连接层 2。通过空间相关注意力，能够提取行人在整个空间内的判别线索，提高行人特征的表示能力。

2.2.2　时间相关注意力

视频中的时序帧包含着很多时间信息，可以通过信息互补，增强行人特征，解决行人遮挡等问题。现有的 Vision-Transformer 方法[27]能够很好地挖掘时序特征，但是其模型庞大的参数量，也增加了训练的难度和时间。因此，文章使用时间相关注意力探索相邻帧的时间相关性。

如图 4（b）所示，时间相关注意力的结构与空间相关注意力相似，同样将 $\boldsymbol{A}_t^{\mathrm{T}}$ 分解为 2 个低维注意力掩码

$$\boldsymbol{A}_t^{\mathrm{T}}=\boldsymbol{S}_t^{\mathrm{T}}\circ\boldsymbol{C}_t^{\mathrm{T}} \tag{11}$$

式中 $\boldsymbol{S}_t^{\mathrm{T}}\in R^{1\times H\times W}$，$\boldsymbol{C}_t^{\mathrm{T}}\in R^{C\times 1\times 1}$ 分别为时间相关注意力分解得到的空间和通道注意掩码。

在空间注意掩码学习中，给定一对相邻帧的特征图 $\boldsymbol{F}_t$，$\boldsymbol{F}_{t+1}$，2 个特征图都视为有 N 个节点的 C 维属性向量 $\boldsymbol{x}_i\in R^C$，$i=1$，…，N。将特征图 $\boldsymbol{F}_t$ 中的一个节点与特征图 $\boldsymbol{F}_{t+1}$ 中所有节点进行相似性

比较，再根据所有双向相似性关系计算得到关系向量：

$$\boldsymbol{r}_{i,j}^{t}=\theta^{\mathrm{T}}\left(\boldsymbol{x}_{t,i}\right)^{\mathrm{T}}\varphi^{\mathrm{T}}\left(\boldsymbol{x}_{t+1,j}\right),\ \boldsymbol{r}_{j,i}^{t+1}=\theta^{\mathrm{T}}\left(\boldsymbol{x}_{t+1,j}\right)^{\mathrm{T}}\varphi^{\mathrm{T}}\left(\boldsymbol{x}_{t,i}\right),\ \boldsymbol{r}_{t,i}^{\mathrm{T}}=\gamma^{\mathrm{T}}\left(\boldsymbol{r}_{i,1}^{t},\cdots,\boldsymbol{r}_{i,N}^{t},\boldsymbol{r}_{1,i}^{t+1},\cdots,\boldsymbol{r}_{N,i}^{t+1}\right) \tag{12}$$

式中：$\boldsymbol{r}_{t,i}^{\mathrm{T}}$为特征图 $\boldsymbol{F}_t$ 中节点 i 的关系向量；θ，φ 和 γ 为 3 个嵌入函数。之后，节点关系值向量 $\boldsymbol{v}_{t,i}^{\mathrm{T}}$由关系向量 $\boldsymbol{r}_{t,i}^{\mathrm{T}}$中对应的嵌入向量堆叠而成

$$\boldsymbol{v}_{t,i}^{\mathrm{T}}=\left[\boldsymbol{r}_{t,i}^{\mathrm{T}},\ \boldsymbol{\beta}\left(\boldsymbol{x}_{t,i}\right)\right] \tag{13}$$

式中 $\boldsymbol{\beta}\left(\boldsymbol{x}_{t,i}\right)$ 的参数与空间相关注意力中所提及的公式一致。最终，用向量 $\boldsymbol{v}_{t,i}^{\mathrm{T}}$更新每个节点位置特征，得到关系值矩阵 $\boldsymbol{V}_t^{\mathrm{T}}$，并通过卷积层和激活函数计算得到空间注意掩码 $\boldsymbol{S}_t^{\mathrm{T}}$

$$\boldsymbol{S}_t^{\mathrm{T}}=\mathrm{Sigmoid}\left(\mathrm{Conv}\left(\boldsymbol{V}_t^{\mathrm{T}}\right)\right) \tag{14}$$

在通道注意掩码学习中，将 $\boldsymbol{F}_t$，$\boldsymbol{F}_{t+1}$2 个特征图送入池化层，得到 2 个特征向量 $\boldsymbol{X}_t^{\mathrm{T}}$，$\boldsymbol{X}_{t+1}^{\mathrm{T}}$，将这 2 个向量堆叠成一个向量，记为 $\boldsymbol{X}_{t,t+1}^{\mathrm{T}}$，之后由 2 个级联的全连接层生成得到通道注意掩码 $\boldsymbol{C}_t^{\mathrm{T}}\in R^{2C}$，公式为

$$\boldsymbol{C}_t^{\mathrm{T}}=\mathrm{Sigmoid}\left(\mathrm{F}_{\mathrm{C2}}\left(\mathrm{F}_{\mathrm{C1}}\left(\boldsymbol{X}_{t,t+1}^{\mathrm{S}}\right)\right)\right) \tag{15}$$

对于计算 $t+1$ 时刻的注意掩码，只需要交换 t 和 $t+1$ 的位置，就可以得到 $t+1$ 时刻的空间相关注意图。

2.3 目标损失函数

与作者先前工作的损失策略不同，文章仅选择了有标签平滑的交叉熵损失（L_{xen}）作为分类损失，学习有利于识别行人的特征表示，同时避免模型过度拟合，使用困难样本挖掘三元组损失（L_{htr}）来优化类间和类内的正负样本分类，提升模型性能。损失公式定义为：

$$L_{\mathrm{xen}}=-\frac{1}{P\cdot K}\sum_{i=1}^{P\cdot K}\log\left(p\left(z_i\mid x_i\right)\right) \tag{16}$$

$$L_{\mathrm{htr}}=\sum_{i=1}^{P}\sum_{a=1}^{K}\left[m+\max_{p=1,\cdots,K}D\left(y_a^i,y_p^i\right)-\min_{\substack{j=1,\cdots,P\\ n=1,\cdots,K\\ j\neq i}}D\left(y_a^i,y_n^j\right)\right]_+ \tag{17}$$

式中：P 和 K 分别表示身份数量与每个身份的采样图像数；y_a^i，y_p^i，y_n^j 分别表示指定样本及其正样本和负样本的特征；D 表示 2 个特征向量的 L_2 范数距离。

另外，在作者先前工作的方法中，输出阶段将图分支和全局分支特征融合成为一个特征，再结合损失函数优化模型。考虑到本文提出的模型是多分支结构，每个分支的优化目标不同，模块参数的更新速率也不一致，使用一个融合特征优化 2 个模块可能达不到最优结果。因此，文章优化了损失策略，首先在训练阶段，将每个分支的输出特征放在一个列表中，再结合损失函数分别更新各自分支模块的参数，在测试阶段，则使用 2 个分支的融合特征和标签计算识别率等评价指标。

联合使用 2 个损失优化本文模型，最终的目标损失函数（L_{tot}）定义为

$$L_{\mathrm{tot}}=L_{\mathrm{xen}}^{\mathrm{glo}}+L_{\mathrm{htr}}^{\mathrm{glo}}+L_{\mathrm{xen}}^{\mathrm{gra}}+L_{\mathrm{htr}}^{\mathrm{gra}} \tag{18}$$

式中：$L_{\mathrm{xen}}^{\mathrm{glo}}$，$L_{\mathrm{htr}}^{\mathrm{glo}}$分别表示对全局分支做交叉熵损失和三元组损失；$L_{\mathrm{xen}}^{\mathrm{gra}}$，$L_{\mathrm{htr}}^{\mathrm{gra}}$分别表示对图分支做交叉熵损失和三元组损失。

3 实验与结果分析

3.1 实验配置

实验在 pytorch1.6.0 的环境下使用 2 个 NVIDIA GTX 2080Ti GPU（11GB 内存）进行。

实验使用的是 2 个主流的基于视频的行人重识别数据集：Mars 数据集和 DukeMTMCM-VideoReID 数据集。采用的评价指标是平均精度均值（mAP）和识别准确率（Rank-k）。

3.2　参数设置

文章实验按照 RRS 策略，从输入视频数据集中选择 8 帧作为图片序列。接着，每一帧图片尺寸都被调整为 256×128（$H\times W$），并通过随机擦除和归一化操作的技巧增强数据。文中使用在 ImageNet[28] 上进行预训练的 ResNet-50 作为骨干网络，按照文献［29］中的设置，将骨干网络的最后一层步长设置为 1，并去掉最后一个空间下采样操作。在训练过程中，采用 Adam 中权重衰减 5×10^{-4} 的策略更新参数，同时设置 3.5×10^{-4} 的初始学习率，并遵循学习率衰减策略。文中模型都训练 500 个轮次。在测试阶段，使用余弦相似度测量训练图像和验证图像距离。

3.3　消融实验分析

为了验证提出方法中各个模块的有效性，文章在 Mars 数据集和 DukeMTMC-VID 数据集上验证不同模型的效果。首先，将作者先前工作提出的方法模型作为本文的 Baseline，具体来说，为了统一框架，利用本文的模型架构和参数设置将先前工作中的图模型和时空注意模块重新进行了实验，并选择交叉熵损失和三元组损失作为目标损失函数。之后，在模型框架中分别加入水平金字塔分割方法和时空相关注意力方法进行实验，在 Mars 数据集上的结果见表 1，其中，py 是水平金字塔分割方法，STA 是时空相关注意力方法，可以看到仅使用其中一个方法时，mAP 和 Rank-1 都略有提升，mAP 分别提高了 0.6% 和 0.9%，Rank-1 分别提升了 0.5%和 0.6%，在 Rank-5，Rank-10 的准确率上变化不大。

联合使用 2 个方法时，精度达到了最好的效果，在 mAP 和 Rank-1 上较 Baseline 分别提高了 1.2%和 1.4%。另外，在 DukeMTMC-VID 数据集上的结果见表 1，使用 2 个方法时，在 mAP 和 Rank-1 上较 Baseline 分别提高了 0.5%和 1.0%。联合使用水平金字塔分割和时空注意力方法，能够逐步聚合局部信息获取全局依赖性，两者结合相辅相成，互相弥补视频中挖掘的潜在信息，同时，图方法能够挖掘更深层的特征，时空注意力方法能够有效减轻遮挡等干扰信息的影响，因此，将 2 个方法结合使用时模型性能更好。

表 1　在 Mars 和 DukeMTMC-VID 数据集上的消融实验

Table 1　Ablation experiments on Mars and DukeMTMC-VID datasets　　%

模型	Mars				DukeMTMC-VID			
	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10
Baseline	84.7	89.6	96.5	97.5	96.6	96.4	99.6	99.9
Baseline+py	85.3	90.1	96.5	97.3	96.7	96.9		
Baseline+STA	85.6	90.2	96.4	97.4	97.1	97.2		
Baseline+py+STA	85.9	91.0	96.7	97.5	97.1	97.4		

3.4　金字塔分割实例分析

文章在 Mars 数据集上进行多次实验，研究了通过改变图像特征分割的金字塔尺度 M 对模型性能的影响。随着金字塔尺度的提升，模型将关注图像中行人划分后更详细的特征。但如果金字塔尺度过多，可能会降低行人的全局信息权重，破坏行人整体特征。另一方面，如果金字塔尺度太少，局部区域的判别特征可能更难提取。因此，选择能够平衡全局和局部特征的金字塔尺度对模型的性能至关重

要，见表 2，可以看到随着尺度（M）降低，mAP 在略微提升，但在 M 为 3 时，Rank-1 效果最佳，因此本文设置了 3 个金字塔尺度。

表 2 金字塔分割实验的结果分析

Table 2 Analysis of results of pyramid segmentation experiment

M	区域/块	mAP/%	Rank-1/%	Rank-5/%	Rank-10/%	Rank-20/%
1	1	86.3	90.7	96.5	97.5	98.3
2	1，2	86.1	90.6	96.5	97.6	98.2
3	1，2，4	85.9	91.0	96.7	97.5	98.3
4	1，2，4，8	85.3	90.6	96.2	97.0	97.7

3.5 视频行人重识别方法比较

在消融实验之后，文章提出的方法将与一些先进视频行人重识别的方法在 Mars 和 DukeMTMC-VID 数据集进行比较，方法包括：AMEM，COSAM[30]，RTF，FGRA，STE-NVAN，TCLNet[31]，STGCN，AFA[32]，MGH[33]，MG-RAFA[34]，AP3D，STMN 等。结果见表 3，与其他先进方法相比，本文的方法取得了较好的结果，在 Mars 数据集上，mAP 达到了 85.9%，与 MG-RAFA 方法相同，可能因为 MG-RAFA 方法采用了多尺度捕捉不同级别语义，因此能获得高 mAP，但本文方法的 Rank-1 达到了 91.0%，较 MG-RAFA 方法高出 2.2%，较 AP3D 方法高出 0.3%，但 AP3D 使用 3D 卷积方法进行时间建模需要更多的计算参数和更高的计算复杂度。在 DukeMTMC-VID 数据集上，本文的方法也达到了 97.1%的 mAP 和 97.4%的 Rank-1。

表 3 在 Mars 和 DukeMTMC-VID 数据集上与最新的方法比较

Table 3 Comparison with the latest methods on the Mars and DukeMTMC-VID datasets %

方法	Mars		DukeMTMC-VID	
	mAP	Rank-1	mAP	Rank-1
AMEM（AAAI20）	79.3	86.7	—	—
COSAM（ICCV19）	79.9	84.9	94.1	95.4
RTF（AAAI20）	85.2	87.1	—	—
FGRA（AAAI20）	81.2	87.3	—	—
STE-NVAN（BMVC19）	81.2	88.9	—	—
TCLNet（ECCV20）	85.1	89.8	96.2	96.9
STGCN（CVPR20）	83.7	89.9	95.7	97.2
AFA（ECCV20）	82.9	90.2	95.4	97.2
MGH（CVPR20）	85.8	90.0	—	—
MG-RAFA（CVPR20）	85.9	88.8	—	—
AP3D（ECCV20）	85.6	90.7	95.6	96.3
STMN（ICCV21）	84.5	90.5	95.9	97.0
Ours	85.9	91.0	97.1	97.4

4 结 论

文章研究并设计了一个基于金字塔分割和注意力机制的视频行人重识别模型，提出了 3 个尺度的水平金字塔分割的方法，将图片特征分别分割成了 1，2，4 块区域，增强图模型对行人整体到局部特

征的识别能力；使用时空相关注意力方法改进时空注意模块，联合局部和全局信息及依赖性，挖掘并互补图片时间和空间相关性信息，学习视频行人的时空特征；更新了模型训练时的参数和架构，并使用交叉熵损失和三元组损失作为目标损失函数。通过在 Mars 和 DukeMTMC-VideoReID 两个数据集上进行的大量实验，验证了本模型的有效性。

参考文献：

[1] NI T G，DING Z Y，CHEN F H，et al. Relative distance metric leaning based on clustering centralization and projection vectors learning for person re-identification[J]. IEEE Access，2018，6：11405-11411.

[2] NI T G，GU X Q，WANG H Y，et al. Discriminative deep transfer metric learning for cross-scenario person re-identification[J]. Journal of Electronic Imaging，2018，27(4)：043026.

[3] WANG H Y，ZHANG W W，SUN J Y，et al. A sparse dimension-reduction based person re-identification algorithm [C]//SPIE Commercial + Scientific Sensing and Imaging. Orlando：SPIE，2018：190-202.

[4] DING Z Y，WANG H Y，CHEN F H，et al. Person re-identification by semi-supervised dictionary rectification learning[C]//SPIE Commercial + Scientific Sensing and Imaging. Orlando：SPIE，2018：172-181.

[5] WANG H Y，WU L Y，CHEN F H，et al. Common-covariance based person re-identification model[J]. Pattern Recognition Letters，2021，146：77-82.

[6] XIAO Y，CAO L，WANG H Y，et al. Unsupervised video-based person re-identification based on the joint global-local metrics[C]//2021 IEEE 7th International Conference on Cloud Computing and Intelligent Systems (CCIS). Xi'an：IEEE，2022：176-182.

[7] 张云鹏，王洪元，张继，等. 近邻中心迭代策略的单标注视频行人重识别[J]. 软件学报，2021，32(12)：4025-4035.

[8] 丁宗元，王洪元，陈付华，等. 基于距离中心化与投影向量学习的行人重识别[J]. 计算机研究与发展，2017，54(8)：1785-1794.

[9] 戴臣超，王洪元，倪彤光，等. 基于深度卷积生成对抗网络和拓展近邻重排序的行人重识别[J]. 计算机研究与发展，2019，56(8)：1632-1641.

[10] 陈莉，王洪元，张云鹏，等. 联合均等采样随机擦除和全局时间特征池化的视频行人重识别方法[J]. 计算机应用，2021，41(1)：164-169.

[11] 徐志晨，王洪元，齐鹏宇，等. 基于图模型与加权损失策略的视频行人重识别研究[J]. 计算机应用研究，2022，39(2)：598-603.

[12] LI J N，ZHANG S L，WANG J D，et al. Global-local temporal representations for video person re-identification [C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul：IEEE，2020：3957-3966.

[13] WU X H，AN W Z，YU S Q，et al. Spatial-temporal graph attention network for video-based gait recognition[M]// Lecture Notes in Computer Science. Cham：Springer International Publishing，2020：274-286.

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

[15] 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). Seattle：IEEE，2020：3286-3296.

[16] 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). Nashville：IEEE，2021：4368-4377.

[17] CHEN L，YANG H，GAO Z Y. Joint attentive spatial-temporal feature aggregation for video-based person re-identification[J]. IEEE Access，2019，7：41230-41240.

[18] ZHU X R, LIU J W, WU H Z, et al. ASTA-net: adaptive spatio-temporal attention network for person re-identification in videos[C]//Proceedings of the 28th ACM International Conference on Multimedia. New York: ACM, 2020: 1706-1715.

[19] ZHANG R M, LI J Y, SUN H B, et al. SCAN: self-and-collaborative attention network for video person re-identification[J]. IEEE Transactions on Image Processing: a Publication of the IEEE Signal Processing Society, 2019, 28(10): 4870-4882.

[20] 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). Montreal: IEEE, 2022: 12006-12015.

[21] 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). Nashville: IEEE, 2021: 2014-2023.

[22] HE K M, ZHANG X Y, REN S Q, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904-1916.

[23] ZHAO H S, SHI J P, QI X J, et al. Pyramid scene parsing network[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017: 6230-6239.

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

[25] LI J N, ZHANG S L, HUANG T J. Multi-scale 3D convolution network for video based person re-identification[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 8618-8625.

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

[27] 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). Montreal: IEEE, 2022: 1470-1481.

[28] DENG J, DONG W, SOCHER R, et al. ImageNet: a large-scale hierarchical image database[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami: IEEE, 2009: 248-255.

[29] LUO H, GU Y Z, LIAO X Y, et al. Bag of tricks and a strong baseline for deep person re-identification[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Long Beach: IEEE, 2020: 1487-1495.

[30] SUBRAMANIAM A, NAMBIAR A, MITTAL A. Co-segmentation inspired attention networks for video-based person re-identification[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul: IEEE, 2020: 562-572.

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

[32] 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? [M]//Computer Vision-ECCV 2020. Cham: Springer International Publishing, 2020: 660-676.

[33] 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). Seattle: IEEE, 2020: 2896-2905.

[34] 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). Seattle: IEEE, 2020: 10404-10413.

（责任编辑：谭晓荷）