中国科学院大学

University of Chinese Academy of Sciences

博士学位论文

基于时空建模的行人重识别研究

作者姓名：侯瑞兵

指导教师：常虹 研究员

学位类别：工学博士

学科专业：计算机应用技术

培养单位：中国科学院计算技术研究所

2022 年 6 月

Person Re-identification Based on Spatial-Temporal Modeling

A dissertation submitted to

University of Chinese Academy of Sciences

in partial fulfillment of the requirement

for the degree of

Doctor of Philosophy

in Computer Science and Technology

By

Hou Ruibing

Supervisor: Professor Chang Hong

Institute of Computing Technology, Chinese Academy of Sciences

June, 2022

中国科学院大学
学位论文原创性声明

本人郑重声明：所呈交的学位论文是本人在导师的指导下独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明或致谢。

作者签名：

日　　期：

中国科学院大学
学位论文授权使用声明

本人完全了解并同意遵守中国科学院有关保存和使用学位论文的规定，即中国科学院有权保留送交学位论文的副本，允许该论文被查阅，可以按照学术研究公开原则和保护知识产权的原则公布该论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存、汇编本学位论文。

涉密及延迟公开的学位论文在解密或延迟期后适用本声明。

作者签名：　　导师签名：

日　　期：　　日　　期：

摘 要

行人重识别是在无交叠视域的摄像头网络中，对同一身份的行人进行跨摄像头的匹配。近年来，行人重识别成为计算机视觉研究的热点，大部分工作集中在基于单帧图像的设置上。相比单帧图像，视频不仅包含行人表观等空间信息，同时也蕴含着行人步态等时序信息，从本质上更容易进行重识别。因此，本文着重研究了基于视频的行人重识别。

视频行人重识别的关键在于时空信息的有效建模。因此，本文的研究目标是如何利用时空信息高效建模完备和鲁棒的行人特征。本文沿着特征更完备，遮挡更鲁棒，计算更高效的思路展开研究，分别提高视频行人重识别模型的准确性、鲁棒性和高效性。本文的主要研究内容及贡献如下：

（1）提出了一种时空交互的单帧判别特征学习方法。针对卷积操作在空间和时间上局部操作导致行人特征缺乏全局信息的问题，设计了一个基于时空交互建模的特征学习方法。该方法通过时空交互操作生成一个时空关系图，有效编码行人部件之间的空间关联和时间关联。基于时空关系图在每帧特征中引入行人全局信息，提高了单帧特征的判别性，进而增强了整合后的视频特征的完备性。在多个重识别数据集上的实验表明，时空交互方法能够显著提高重识别准确性，验证了全局建模的有效性。

（2）提出了一种时序互补的多帧完整特征学习方法。沿着上一个工作，进一步探索了视频“特征完备性”的问题。针对上一个工作中，由于无差别处理每一个视频帧而导致的相邻帧特征冗余的问题，设计了一个基于时序互补建模的特征学习方法。时序互补建模的核心思想是显式约束连续帧提取互补特征，使整合的视频特征能够覆盖一个更完整的行人身体区域，从而提高视频特征的完整性。实验证明所提方法能够有效减小连续帧的特征冗余，提高特征的完备性，进一步提升行人重识别的准确性。

（3）提出了一种遮挡鲁棒的时空补全方法。针对遮挡条件下的视频行人重识别问题，受到行人视频具有丰富的时空结构性的启发，先后提出了图像和特征层面的时空遮挡补全方法。图像级别的时空补全方法采用从粗到细的思想，渐进式地利用空间和时间信息复原遮挡区域的像素。为了避免昂贵的图像生成，

进一步设计了一个特征级别的时空补全模块，直接修复遮挡区域的特征表示。相比于图像补全，特征补全方法能够根据行人重识别的反馈信号直接进行优化，从而充分提高重识别模型对遮挡的鲁棒性。实验表明，时空补全方法在真实遮挡的行人重识别数据集上能够取得明显的性能提升，验证了所提方法对遮挡的鲁棒性。

（4）提出了一种高效的时空特征学习框架。针对上述工作以高分辨率处理多帧导致的效率较低的问题，设计了一个基于双分支结构的高效时空建模框架。在空间特征学习方面，以相同的时间间隔降低视频帧的分辨率，在降低计算开销的同时提取到多尺度表观特征。在时序特征学习方面，设计了一个轻量的时序核选择模块。通过自适应捕捉视频的短期和长期时序关系，提取到多尺度动作特征。实验表明，所提框架可以与现有视频重识别工作有效整合，在维持重识别的准确性和鲁棒性的同时，降低了大约50%的计算开销，提高了重识别模型的效率。

综上所述，本文针对视频行人重识别任务开展了深入的研究，从时空建模出发提出了多种创新模型。大量实验结果表明，本文提出的方法可以有效提升视频行人重识别的准确性、鲁棒性和高效性。

关键词： 行人重识别，深度学习，视频，时空建模

Abstract

This thesis studies person re-identification (reID), which aims to match pedestrians with the same identity across non-overlapping cameras. It is a basic technique for intelligent surveillance system. Recently, person reID has become a research hotspot in computer vision. Most works focus on image setting. Compared to single-frame image, video contains richer spatial appearance information and extra temporal motion information, which is inherently easier to reID. Therefore, this thesis mainly studies video person reID.

The key to video person reID is the effective spatial-temporal modeling. This thesis aims to study how to efficiently model spatial-temporal information to learn a complete and robust person feature. The research is carried out along the idea of more complete features, more robust to occlusion and more efficient calculation, to improve accuracy, robustness and efficiency of video reID model respectively. The main contributions of this thesis are summarized as follows:

(1) A spatial-temporal interaction approach to learn discriminative feature for each video frame. Convolutional operations typically process a local neighborhood in space and time, thus is limited to local context modeling. To deal with this problem, we design a spatial-temporal interaction module to perform global context modeling. This approach first generates a spatial-temporal relation map to effectively encode the spatial and temporal relations across disjoint and distant body parts. Based on the relation map, the global spatial-temporal contexts can be incorporated into each frame, which improves the discriminability of single-frame feature and further manifests the final video features. Extensive experiments on multiple reID benchmarks show the proposed method can effectively model long-range spatial and temporal relations, enhance the global modeling ability, and significantly improve the reID accuracy.

(2) A temporal complementary learning approach to learn integral video features. This work further explores the "feature completeness" based on the first work. The first work performs a same operation on each frame, which typically produces highly redun-

dant features for consecutive frames. To deal with this problem, we design a temporal complementary approach. The key idea is to explicitly constrain different consecutive frames to mine complementary parts, so that the video features can cover more integral person characteristic. Experiments show that the proposed method can effectively reduce the feature redundancy of consecutive frames, enhance the completeness of video feature, and further improve the reID accuracy.

(3) A occlusion-robust spatial-temporal completion approach. Inspired by the spatial-temporal structure of pedestrian sequences, we successively propose image-level and feature-level spatial-temporal completion approach for occluded video person reID. The image-level spatial-temporal completion approach adopts a coarse-to-fine strategy, which utilizes spatial and temporal clues to gradually restore the occluded pixels. To avoid the expensive image generation, we further propose a feature-level spatial-temporal completion approach, which directly repairs the feature representation of occluded regions. In this way, the feedback signals from reID task can supervise the occlusion completion to fully improve the reID performance. Experiments show that the proposed approach performs favorably against state-of-the-arts on occluded datasets, which validates the occlusion robustness of the proposed spatial-temporal completion mechanism.

(4) An efficient spatial-temporal feature learning framework. Above works process each frame at a high input resolution with low computational efficiency. To this end, we design an efficient spatial-temporal modeling framework based on a bilateral structure. For spatial feature learning, by reducing the input resolution of some video frames at time intervals, the computational overhead is effectively reduced and a multi-scale appearance feature can be obtained. For temporal feature modeling, we design a lightweight temporal kernel selection block, which adaptively varies the scale of temporal modeling on the properties of input videos, thereby learning a multi-scale motion feature. Experiments show that the proposed framework can effectively integrate with existing video reID works, which maintains accuracy and robustness while improving efficiency by reducing about 50% computations.

In conclusion, this thesis studies video person reID, and proposes several ap-

proaches based on spatial-temporal modeling. Experimental results on various datasets show that the proposed approaches improve the accuracy, robustness, and efficiency of video person reID.

Keywords: person re-identification, deep learning, video, spatial-temporal modeling

目 录

图形列表

表格列表

第1章　绪论

1.1　课题研究背景和意义

随着社会和经济的发展，越来越多的监控摄像头部署到公共场所，带来了海量的监控视频数据。传统的方法主要依靠人工完成对监控视频的分析，十分消耗人力和时间。为了处理海量视频数据，智能视频监控应运而生。智能视频监控通过计算机视觉等领域的相关技术，自动对视频内容进行分析，快速实现监控视频中的目标检测、识别、跟踪、以及动作和行为分析等。其中一个关键技术是利用视频内容进行身份鉴定，即行人重识别（Person re-identification，reID）技术。

行人重识别是在跨摄像头环境下识别特定行人的技术。行人重识别在智能监控中有非常重要的应用价值。比如，在刑侦方面，根据监控视频自动检索嫌犯位置，获得嫌犯的移动轨迹，提高警方的办案效率；在日常生活方面，帮助搜索失踪人员、找回迷路的儿童和老人、帮助群众找回遗失物品等等。此外，行人重识别技术涉及到图像处理与分析、机器学习、模式识别等多种学科，因此行人重识别技术的发展可以进一步促进相关学科的发展。总结来说，行人重识别技术在学术研究和实际应用中都有重要的价值，使之成为当前计算机视觉研究的热点问题。

行人重识别技术经过几十年的快速发展已经取得了显著的进展，在简单场景下已经取得不错的性能 [1]。然而在复杂场景下，由于受到姿态、遮挡、图像分辨率低等多种因素的影响，行人重识别依旧面临巨大挑战。而相比于行人图像，视频数据具有天然的优势。（1）由于监控行人以视频的形式呈现，视频数据更贴近实际应用场景。（2）视频数据具有更广阔的应用场景。图像行人重识别只能利用单帧图像的视觉线索，因此对图像质量要求较高，导致相机布置与使用场景比较受限。而视频数据可以利用多帧的视觉线索，因此包含更丰富的信息，对图像质量要求相对较低，可以应用到更复杂的监控场景。（3）视频中不同帧的多样性提供了更为丰富的行人外观信息。图像行人重识别只能从单帧图像中获取有限的行人信息，在解决单帧图像的遮挡等问题上效果有限。相比之下，视频序列的不同帧具有不同的姿态和视角，能够提供更丰富的行人外观

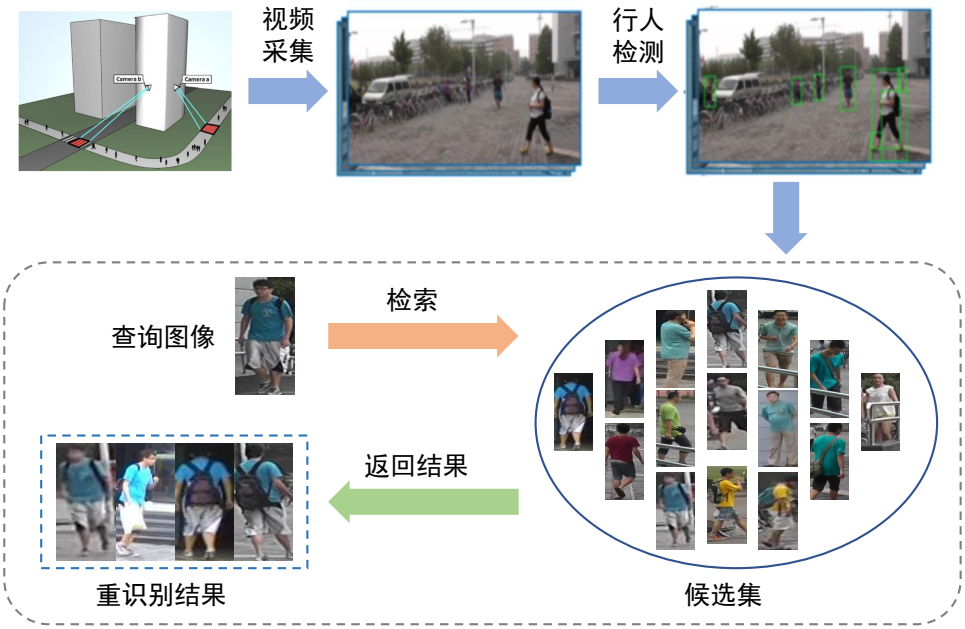


图 1.1 广义行人重识别系统的完整流程图

信息，有利于建模更鲁棒的外观表示。（4）视频数据提供了时序信息。相对于单帧图像，视频不仅包含外观等空间信息，同时也具有步态等时序信息。视频的时序信息能够帮助区分表观相似的不同行人。由于以上优势，视频行人重识别技术可以更好地工作在复杂场景，受到了研究者越来越多的关注。

1.2 行人重识别问题描述

广义上，行人重识别研究的目的是赋予计算机与人类视觉系统相同的行人识别工作。一个广义的行人重识别系统主要包括视频采集，行人检测以及行人重识别三个过程，如图1.1所示。下面分别予以介绍。

视频采集 行人重识别系统首先要做的是对行人视频进行采集，通过摄像机、摄像头、数码相机等都可以采集到行人视频。

行人检测 采集到视频之后，便需要确定行人所在的位置和大小进而分离出行人区域，即行人检测。并将所有行人检测结果汇聚在一起，构成候选集（gallery set），同时人工划分出一部分数据作为查询集（query set）。

行人重识别 行人检测之后，就可以利用检测区域内的行人信息进行下一步的识别了，即行人重识别。行人重识别算法的目标是给定查询集（query）中某个行人的图像或者视频，在无重叠视域的摄像头下的候选集（gallery）中识别

出身份相同的行人。行人重识别算法一般包括图像/视频预处理、特征提取、距离度量以及候选结果排序等环节。目前，行人重识别研究主要针对上述的一个或多个环节进行改进，提高行人重识别算法的精度和速度。

本文关注于广义行人重识别系统中的基于视频的行人重识别算法部分。

1.3　行人重识别的主要挑战

从问题的提出到现在，行人重识别已经经过了几十年的研究和发展，在简单受控的场景下已经取得不错的性能并成功应用到现实生活中。但是在复杂情况下，还远远不能满足智能视频监控的要求。如何在复杂场景下准确而快速实现行人重识别，仍是一个尚未完全解决的问题。这是由行人本身的特性和复杂的环境因素决定的。目前，行人重识别技术主要面临以下三方面挑战：

类间差异小–不同行人着装相似　虽然每个人的体型、身高、衣物等都是不同的，但却都具有相似的结构和外形，比如所有人的头、胳膊、腿等的相对位置和外形都是大致相似的。人类的身体结构特性导致不同个体之间的差异较小，减弱了不同人体之间的区分程度。更为重要的是，一些监控场景中，大量的不同行人会穿相似款式或颜色的衣服，比如校园中大多数学生会穿统一的校服，导致不同身份行人的表观可能高度相似。而为了使单个摄像头覆盖大范围监控区域，监控摄像头所采集到的行人图像分辨率较低。在低清晰度的行人图像中，几乎无法通过脸部信息辨别行人身份，使得现阶段算法主要依据行人穿着等特征做决策。如图1.2所示，这些着装相似的不同行人在低清晰度图像上具有极强的相似性，增大了行人重识别任务的挑战性。

类内差异大–遮挡　由于一些不可避免的因素影响，同一行人在不同摄像头下的图像或者视频可能存在较大的表观变化，造成类内差异大问题。遮挡是导致类内差异大的一个极具挑战性的因素。在监控场景中，尤其是人流密集的场所，行人与行人之间以及行人与监控环境中存在的其他物体之间存在大量的遮挡。如图1.3所示，遮挡物覆盖了行人腿部等身体部位，导致目标行人的可视区域的减少以及对应特征的丢失。遮挡物的不同加大了同一个行人在不同摄像头下的表观差异，增加了行人重识别任务的难度。

视频计算效率较低　基于图像的行人重识别只能从单个图像中获取有限信息，在解决不同行人表观相似、遮挡等问题上效果有限。相比之下，视频包含

图 1.2 行人重识别任务的难点：不同行人穿着相似导致的类间差异小

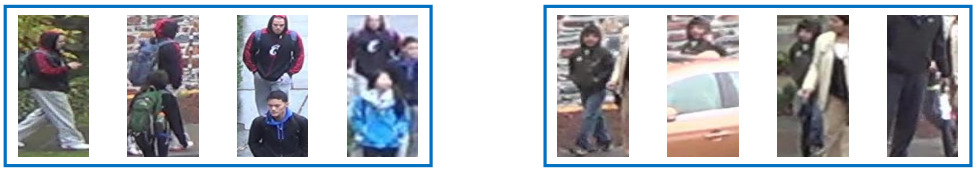


图 1.3 行人重识别任务的难点：遮挡因素导致的类内差异大。每个蓝色框的四张图像属于同一个行人

更丰富的时空信息，从本质上更容易进行重识别，比如可以通过步态等时序信息帮助区分着装相似的不同行人。然而，在视频行人重识别中，由于每次要处理多张图像，视频建模的计算效率较低，如何提高视频建模的计算效率也是实际应用中一个必须考虑的问题。

1.4 行人重识别研究现状

在2006年的国际计算机视觉与模式识别会议上，行人重识别（Person Re-identifacation）的概念 [2] 被首次提出，随后与之相关的研究不断涌现出来。2010年Farenzena等人 [3] 和Bazzani等人 [4] 提出了视频行人重识别（Video Person Re-identification）任务。2014年，随着大规模数据集的出现和具有强大并行计算能力的图像处理单元的使用，以及AlexNet [5] 在图像分类任务上的成功所引领的神经网络的复兴，行人重识别进入了深度学习时代 [6–12]。当前，以卷积神经网络（Convolutional Neural Network，CNN）为代表的深度网络已经成为行人重识别方法的主流，本文围绕基于深度学习的行人重识别方法进行讨论和研究。

如图1.4所示，基于深度学习的行人重识别方法可以整合到一个框架中：给

定输入图像或视频，首先使用深度卷积神经网络提取其特征表示，然后通过分类（Classification）[13–15] 或者度量学习（Metric Learning） [16–21] 的损失进行训练，形成一个端到端的学习框架。现有研究主要集中在特征提取和损失函数设计两个方面。本文侧重于研究特征提取方面，主要探讨的是行人重识别中行人特征表达的关键技术。行人重识别任务根据研究对象的不同可以分为基于图像的研究和基于视频的研究。在本章的后续小结中，将分别对这两类工作进行综述。

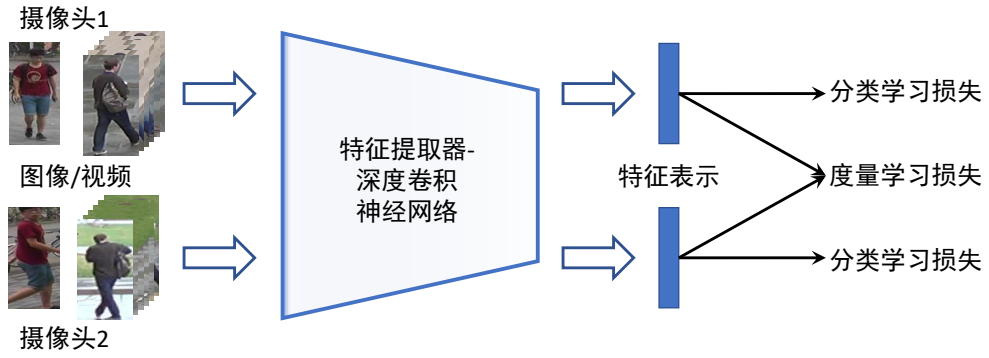


图 1.4 基于深度学习的行人重识别方法框架

1.4.1　基于图像的行人重识别

行人重识别作为一个细粒度检索任务，主要存在类内差异大和类间差异小两个问题。针对类内差异大问题，一类经典方法是提取能够反映行人身份并且对各种变化因素不敏感的特征。从早期的姿态鲁棒特征到最近的遮挡鲁棒特征，研究者们不断改进对行人的特征描述，试图提取能够更好适应行人表观变化的特征表示；针对类间差异小问题，主流方法是提取能够覆盖行人整体区域的特征。下面分别对这三种算法类型进行简单的梳理。

类内差异大–姿态问题　在特征提取器的构建上，早期研究主要解决行人的姿态变化问题 [22–24]。 Hou等人 [25] 指出卷积单元固定的几何结构限制了特征提取器对姿态变化的鲁棒性。基于这个分析，Hou等人 [25] 提出了一个基于自注意力机制 [26] 的交互整合方法（Interaction-Aggregation，IA），通过自适应确定不同部件感受野的大小和形状，得到对姿态等非刚性变化鲁棒的特征提取器。Xia等人 [9] 将IA [25] 拓展为二阶形式，进一步提高模型对姿态变化的鲁棒性。目前，基于自注意力的行人重识别方法 [9, 27, 28] 在具有较大姿态变化的Market1501数据集 [29] 上能达到95%以上的准确率，说明姿态变化已经不再

是行人重识别的瓶颈。

类内差异大–遮挡问题 Miao等人 [22] 较为系统地定义了遮挡行人重识别问题。遮挡行人重识别的特点是使用包含遮挡物的半身图与行人全身图进行匹配。 Miao等人 [22] 提出了一个针对遮挡问题的常用框架：首先使用离线的姿态估计网络[30, 31]对行人进行部件区域划分，然后提取每个部件的特征并判断该部件是否发生遮挡，最后根据图像对间共有区域的特征计算相似性。在这个框架上，文献 [32, 33] 将离线的姿态估计网络改为在线训练，使其更好适用当前数据。为了避免使用计算昂贵的姿态估计网络，文献 [34–37] 使用轻量级的空间注意力定位可见的身体部件区域。以上方法的主要思路是让模型能够抑制对遮挡区域的响应，会造成行人外观信息的损失。Li等人 [38] 使用Transformer [26] 挖掘部件之间的关联，改善遮挡区域的特征表示。

类间差异小–不同行人着装相似问题 针对行人重识别的类间差异小问题，现有模型 [24, 39] 致力于挖掘更多的细粒度视觉线索：比如鞋子，背包和帽子等。这类工作借助外部的姿态估计网络或者行人分割网络定位多个判别的部件区域，然后利用多分支网络从多个局部区域中提取细粒度特征。尽管这类方法能够关注到更多的行人表观信息，但是存在以下几个问题：（1）姿态估计模型，行人分割模型以及多分支结构都会大幅增加模型的计算开销；（2）姿态估计模型和行人分割模型无法定位到背包、帽子等属性信息；（3）姿态估计模型和行人分割模型难以避免存在误差，这些误差会传入到行人重识别模型中损害局部特征的判别性。

1.4.2 基于视频的行人重识别

经过近几年的发展，图像行人重识别已经得到了深入的研究。但是由于图像中的信息有限，在严重遮挡和大量不同行人着装相似的场景上，目前图像行人重识别的方法仍然无法取得令人满意的结果。比如，在具有严重遮挡的图像数据集Occluded-DukeMTMC [40] 上，目前最好的方法PAT（Part-Aware Transformer） [38] 仅能达到50%左右的准确率；在存在大量着装相似的图像数据集MSMT17 [41] 上，目前最好的方法ABD-Net（Attentive But Diverse Network） [42] 的准确率大约为60%左右。从这些结果中可以看到，图像行人重识别方法在解决遮挡，不同行人着装相似的问题上效果有限。相比之下，视频数据包含更多可利用信息，包括行人的外观、姿态和运动等信息，从本质上更容易识别行

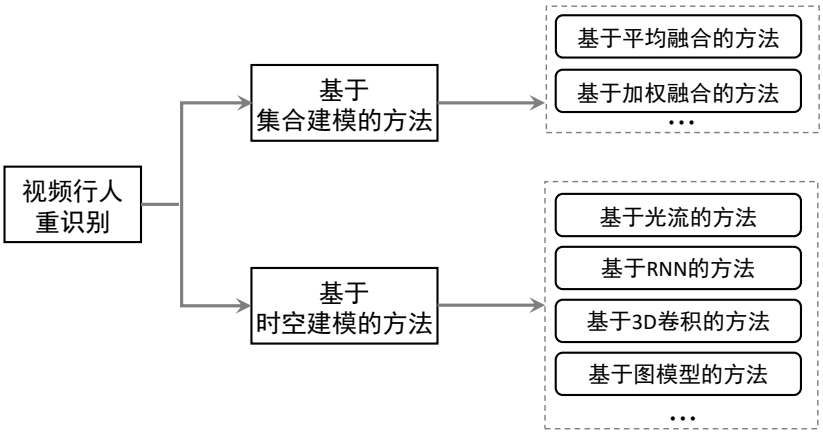


图 1.5 基于深度学习的视频行人重识别方法分类

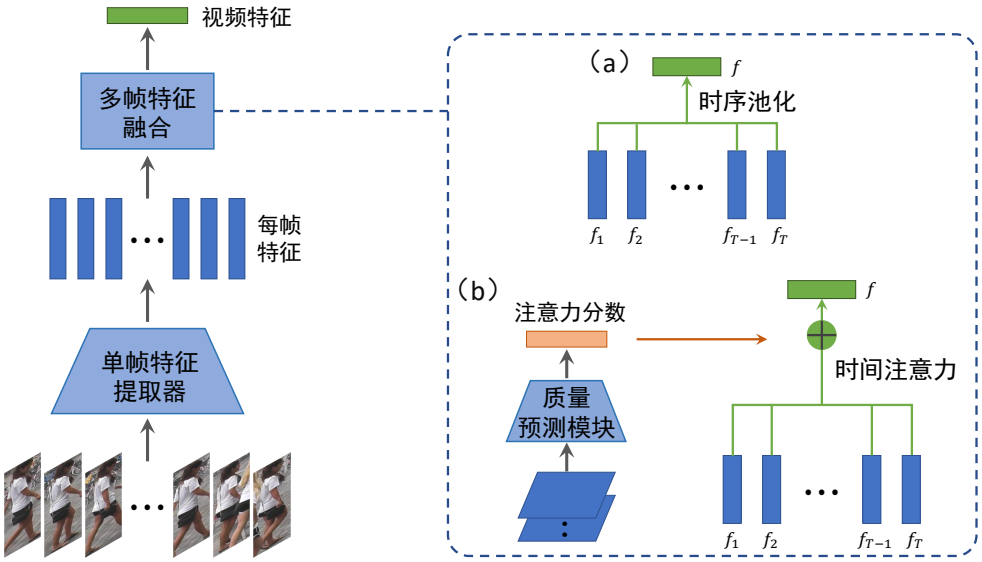


图 1.6 视频行人重识别中基于集合建模的方法框架

人。因此，基于视频的研究显得尤为重要。

如图 1.5所示，根据是否利用视频的时序信息，视频行人重识别方法可以分为基于集合建模的方法 [43–47] 和基于时空建模的方法 [48–54]。下面分别予以介绍。

1.4.2.1　基于集合建模的方法

基于集合建模的方法是早期视频行人重识别的一类典型方法。这类方法把视频的每一帧看做独立的图像，用无序的图像集合表示视频序列。基于集合建模的方法一般可以整合到一个网络框架中。如图1.6所示，给定输入的视频序列，首先利用卷积网络独立提取每一帧图像级别的特征，然后融合所有帧的特征得到一个视频级别的特征表示。各种集合建模方法的不同主要体现在多帧融合策

略的设计上。其中两个代表性方式是时序池化（Temporal Pooling） [43] 和时间注意力（Temporal Attention） [44–47]。如图1.6(a)所示，时序池化融合沿着时间维度对所有帧的特征执行平均池化或者最大池化得到视频特征。但是由于遮挡、光照等因素都可能造成图像质量不佳，一个视频序列中往往并不是所有帧都具有较高的质量，因此保留更多高质量帧的特征尤为重要。如图1.6(b)所示，时间注意力对每一个视频帧进行质量评估，并以质量分数为权重对所有帧的特征进行加权求和。基于集合建模的方法简单有效，但是忽视了视频的时序信息。

1.4.2.2 基于时空建模的方法

基于视频时空建模的方法保留了视频的时序关系。合理利用时序关系能够增强行人的外观表示，并且反映行人的运动特性。因此，越来越多的研究者将研究重点转移到如何更有效建模视频的时空信息。主流方法使用光流、循环网络、三维卷积、图网络等建模时空信息，下面分别予以介绍。

- 基于光流的时空建模方法

早期的工作使用光流（Optical Flow）编码相邻帧间的运动。这类方法的研究集中在如何整合单帧包含的表观信息以及光流包含的运动信息。McLaughlin等人 [48] 提出在输入层面上融合图像和光流信息，将原始图像和对应光流结合作为网络新的输入。 Liu等人 [55] 提出在特征层面上融合图像和光流信息，该方法使用两个子网络分别提取图像的表观特征和光流的动作特征，然后融合这两种特征得到时空表示。尽管光流可以有效提取运动特征，但是手动提取光流的复杂度较高，难以部署常见的要求实时的场景。此外，光流只能编码相邻帧的短期运动，无法建模序列的长期时序关系。

- 基于循环网络的时空建模方法

循环网络（Recurrent Neural Network, RNN）[56] 以及其变体长短时记忆网络（Long Sort-term Memory，LSTM）[57] 常用于长期时序建模。基于RNN的方法 [48, 58–61] 一般将CNN提取的帧级别特征输入到RNN中，进而对视频进行时序建模。尽管早期工作 [48, 58–61] 证明了循环网络在视频行人重识别任务的有效性，但是这些工作使用较为简单的卷积网络（比如四层卷积网络）作为空间特征提取器。Gao等人 [62] 发现当使用一个较深的卷积网络（比如ResNet50）提取空间特征时，循环网络的表现甚至低于时序平均池化操作。Zhang等人 [63]

尝试打乱RNN输入帧的时间顺序，发现打乱后的结果甚至高于正常输入的结果。这些实验都从一定程度上说明当使用一个判别性较强的卷积特征作为RNN的输入时，RNN无法有效建模序列信息。因此近两年很少有工作使用基于循环网络的时空建模方法。

- 基于三维卷积的时空建模方法

三维卷积（3D Convolution）是二维卷积在三维空间的拓展，最早应用在视频识别任务中 [64]。相比于二维卷积，三维卷积增加了一个深度通道，可以沿着空间和时间维度操作联合提取视频的时空特征。但是三维卷积引入了一个新的维度，导致三维卷积网络参数量巨大，难以训练。P3D（Pseudo-3D）[65] 将三维卷积核分解成一个二维的空间卷积核和一个一维的时间卷积核，有效减小了三维卷积的计算开销和参数量。 Liao等人 [49] 将P3D应用到视频行人重识别任务中。 Gu等人 [51] 提出了表观保留三维卷积（Appearance-Preserving 3D，AP3D），首先将相邻帧进行空间对齐，之后再将对齐后的特征进行三维卷积。Li等人 [52] 提出了M3D（Multi-scale 3D），将P3D中的一个时序卷积核替换为多个不同尺度的时序空洞卷积核，同时建模行人序列的短期和长期的时序关系。M3D能够同时建模短期和长期的时序关系，具有更强的时序建模能力。

- 基于图网络的时空建模方法

最近的一些研究 [53, 66–71] 利用图网络 [72] 建模行人序列的时空信息。Yan等人 [66] 构建了一个多尺度超图（Multi-Granular Hypergraph，MGH），首先多粒度水平分割每帧的卷积特征图得到多个区域节点，之后连接每个粒度的所有节点构建超图，最后基于构建的超图捕捉序列的时空依赖。文献 [53, 67] 构建了一个空间图网络和一个时序图网络，分别建模视频的空间关系和时序关系，有效降低了时空建模的难度。更进一步，He等人 [69] 在图网络中引入时空位置信息。 Wang等人 [70] 设计了一个金字塔型时空整合框架（Pyramid Spatial-Temporal Aggregation，PSTA），同时兼顾短期和长期的时序关联。

1.5　行人重识别技术的性能评测

行人重识别技术的发展伴随着一些经典数据集的发布，以及公开的评价体系的建立。在行人重识别发展的早期阶段，由于资源限制，发布的大多是小规

模且较为简单的行人数据集。随着行人重识别技术的发展，小规模行人数据集已经无法满足重识别方法的评测要求，一些较大规模的数据集逐渐受到关注。下面分别对行人重识别的主流数据集和衡量指标进行介绍。

1.5.1 行人重识别数据集

表1.1按照数据集的数据形式、数据规模和发布时间等梳理了行人重识别领域的公开及常用的数据集。下面按照发布的时间分别对主流的图像和视频行人重识别数据集进行介绍。

表 1.1 典型行人重识别数据集

数据集	ID数	图片数	视频数	相机数	检测器	发布时间
图像数据集						
CUHK03 [73]	1,467	13,164	-	2	手动+DPM	2014
Market-1501 [29]	1,501	32,668	-	6	手动+DPM	2015
Duke [74]	1,812	36,441	-	8	手动	2017
MSMT17 [41]	4,101	126,441	-	15	Faster RCNN	2018
Occluded-Duke [22]	1,812	35,484	-	8	手动	2019
视频数据集						
iLIDS-VID [75]	300	42,460	600	2	手动	2014
MARS [43]	1,261	1,067,516	20,715	6	DPM+GMMCP	2016
Duke-Video [76]	1,812	815,420	4,832	8	手动	2018
LS-VID [77]	3,772	2,982,685	14,943	15	Faster RCNN	2019
Occluded-Duke-Video [78]	1,812	573,457	4,338	8	手动	2021

- CUHK03行人图像数据集 [73]

由香港大学（The University of Hong Kong，HKU）构建。该数据集提供了手动检测和可变形部件模型检测（Deformable Part Mode，DPM） [79] 检测两个版本。其中DPM检测数据集包含检测误差，更接近实际场景需求。该数据集来自2个摄像头，共包含1,467个行人的13,164张图像。

- Market-1501行人图像数据集 [29]

由清华大学（Tsinghua University）创建。该数据集包含1501个行人，来自6个不同摄像头。其中训练集和测试集分别包含751和750个行人，12,936和19,732张

图像。所有图像均由DPM自动检测得到。迄今为止，Market-1501数据集仍是行人重识别中广泛应用的数据集之一。

- DukeMTMC-reID行人图像数据集 [74]

由杜克大学（Duke University）构建，该数据集包含1,812个行人，来自8个不同摄像头。其中训练集包含702个行人，共16,522张图像，测试集包含1,110个行人，共17,661张图像。所有图像均由手动检测得到。

- MSMT17行人图像数据集 [41]

由北京大学（Peking University）发布。它采用15个摄像头，包括12个户外摄像头和3个室内摄像头，覆盖了多个复杂场景。该数据集涵盖了早上、中午和下午三个时间段，具有复杂的光照变化，包含4,101个行人和126,441张图像。所有图像均由Faster RCNN [80] 检测得到。MSMT17是目前最大规模的图像行人重识别数据集之一。

- Occluded-DukeMTMC行人图像数据集 [22]

2019年由百度研究院（Baidu Research）重新组织了DukeMTMC-reID数据集所创建。该数据集包含1,812个行人和35,484张图像，具有大量复杂的遮挡模式，是目前遮挡场景下最大的图像行人重识别数据集之一。

- iLIDS-VID行人视频数据集 [75]

由清华大学（Tsinghua University）根据i-LIDS Multiple-Camera Tracking Scenario [81] 数据集构建。iLIDS-VID来自2个摄像头，包含300个行人，其中每个行人包含2个序列。该数据集涵盖了大量严重的遮挡，是一个极具挑战的视频数据集。

- MARS行人视频数据集 [43]

由清华大学（Tsinghua University）创建。MARS是Market-1501数据集的扩展，包含了每个行人图像的整个跟踪序列。MARS提供了1,261个行人的20,715个序列，平均每个序列有50帧。由于MARS的序列是由早期的DPM [79] 检测器和GMMCP [82] 跟踪算法得到，序列的时序连续性较差。到目前为止，MARS数据集仍然是规模最大、难度最高的数据集之一。

- DukeMTMC-Video行人视频数据集 [76]

由华南理工大学南科大-悉尼科技大学联合中心（SUSTech-UTS Joint Centre of CIS, Southern China University of Science and Technology）根据DukeMTMC数据集构建，是DukeMTMC数据集的视频版本。DukeMTMC-Video包含1,812个行人和4,832个图像序列，平均每个序列有167帧。该数据集不同序列的帧数变化非常大，最短的序列仅仅有1帧，最长的序列有9,324帧。不同序列长短不一，加大了该数据集的视频建模的难度。

- LS-VID行人视频数据集 [77]

由北京大学（Peking University）构建，是MSMT17数据集的扩展。LS-VID包含3,772个行人，14,943个图像序列，平均每个序列有200帧。所有图像采用Faster RCNN [80] 检测输出，具有较为精确的检测结果。该数据集覆盖了多个复杂场景，并包含了早上，中午和晚上不同的时间段，涵盖了光照、姿态、图像质量等很多方面的变化，非常具有挑战性。LS-VID是目前最大规模的视频行人重识别数据集之一。

- Occluded-DukeMTMC-Video行人视频数据集 [78]

由中科院计算所（Institute of Computing Technology of the Chinese Academy of Sciences，CAS）基于DukeMTMC-Video数据集重新组织构建。该数据集包含1,812个行人，4,338个序列，平均每个序列有132帧，是目前遮挡场景下最大的视频行人重识别数据集。

1.5.2 行人重识别评价指标

行人重识别任务广泛采用两种评价指标，包括累计匹配曲线（Cumulative Matching Characteristic，CMC）[83] 和均值平均精度（Mean Average Precision，mAP）[84]。

累计匹配曲线（Cumulative Matching Characteristic，CMC）是检索任务中常用的一个指标，在行人重识别、图像检索、人脸识别等领域被广泛使用。给定查询图像，对候选集的图像根据特征距离排序，目标行人的排序位置可以指示重识别模型的效果。在跨摄像头的行人重识别任务上，需要排除候选集中与查询图像处于同一个摄像头的图像，防止其参与排序。具体来说，假定查询集中

图像的个数为N，$index_i$表示候选集中与第i个查询图像为相同行人的最靠前的排序结果。最终top-k准确度$A\,(\text{top-k})$计算为：

$$A\,(\text{top-k}) = \frac{1}{N}\sum_{i=1}^{N}\begin{cases} 1 & index_i \le k \\ 0 & index_i > k \end{cases} \tag{1.1}$$

最终以k为横坐标，准确度$A\,(\text{top-k})$为纵坐标绘制CMC曲线。在实际使用中，比较常用的是top-1、top-5和top-10，代表了在前1、 5、 10张候选图像中出现目标行人的概率。

但是，当图像库中存在一幅以上的查询图像的正确匹配时，CMC曲线不足以反映方法召回率方面的性能。针对这个问题，Zheng等 [29] 提出使用均值平均精度（Mean Average Precision，mAP）作为补充的评价指标。mAP的计算需要以下三个步骤。首先，计算精准率（Precision）。对于每张查询图像q_i，返回候选集的排序结果。假定前n个排序结果中与q_i为相同行人的数目为$c\,(n, q_i)$，精准率$P\left(n, q_j\right)$计算为：

$$P\left(n, q_j\right) = \frac{c\,(n, q_i)}{n}. \tag{1.2}$$

然后，计算平均精准率（Average Precision，AP）。假定查询图像q_i在候选集中有M个正样本，记录所有M个正样本的排序结果$(i_1, i_2, \ldots, i_M)$，计算它们的平均精准率：

$$AP\left(q_j\right) = \frac{\sum_{i\in\{i_1,i_2,\ldots,i_M\}} P\left(i, q_j\right)}{M}. \tag{1.3}$$

最后，均值平均精度（Mean Average Precision，mAP）通过所有查询图像AP的均值获得：

$$mAP = \frac{\sum_{j=1}^{N} AP\left(q_j\right)}{N}. \tag{1.4}$$

mAP综合考虑了准确率和召回率，可以作为CMC的补充，提供更为全面的性能评价。

1.6　本文拟解决问题与主要贡献

1.6.1　拟解决问题

行人重识别作为一个细粒度检索任务，存在两个本质难点，类内差异大和类间差异小：由于姿态变化、遮挡等影响，同一个行人在不同摄像头下的图像可能存在较大的外观差异；同时，真实场景往往存在大量着装相似的不同行人，

(a) 视频序列的连续三帧　　(b) 目前方法的特征激活图

图 1.7 行人序列的特征激活图示例

图 1.8 遮挡场景下的行人序列示例

导致不同行人的外观差异可能非常小。由于图像的信息有限，基于图像的行人重识别在解决上述问题上效果有限。

相比于图像，视频具有更丰富的外观信息以及运动信息，从本质上更容易进行重识别。因此，本文侧重研究基于视频的行人重识别。从前述章节的介绍可以看出，已有视频行人重识别方法存在以下三方面主要问题，使其依旧受限于解决类间相似和类内差异问题。

（1）常规的时空建模方法提取的特征完备性较差

现有时空建模方法通常使用卷积网络无差别处理每一帧，然后再利用RNN等进行时序建模。当卷积网络已经关注到一个可以识别目标行人的局部区域时，便不会再去关注其他身体区域 [39]。因此卷积网络通常只关注视频帧的一个局部区域。其次，由于连续帧的高度相似性，无差别处理所有帧导致连续帧特征的高度相似和冗余，这些高度相似的特征趋向于关注到同一个局部区域。一个示例如图1.7所示。因此，常规时空建模方法不能得到一个完备的行人视频特征，依旧难以区分类间相似的不同行人。

（2）针对遮挡的时间注意力方法提取的特征鲁棒性较差

针对遮挡问题，目前方法采用的策略是利用时间注意力对遮挡帧赋予一个低的权重。然而，如图1.8所示，遮挡帧中仍然存在具有判别线索的可视区域，并且维持了一个序列的完整时序信息。因此，时间注意力方法会造成行人外观信息的损失以及时序信息的中断，加大了同一个行人的特征差异，导致仍然无

法得到一个遮挡鲁棒的行人特征表示。

（3）计算效率较低

由于要以一个高分辨率同时处理序列的多帧图像，大多数视频行人重识别方法具有较低的计算效率。如何提高视频建模的计算效率也是视频行人重识别亟需解决的一个问题。

1.6.2　研究内容与主要贡献

通过对行人重识别国内外研究现状的回顾、归纳和总结，以及对目前研究存在问题的分析，本文确立了如下的研究目标：面向视频行人重识别任务，研究如何利用时空信息高效建模完备和鲁棒的行人特征，提高行人重识别系统的准确性、鲁棒性和高效性。如图1.9所示，本文工作沿着特征更完备（系统准确性），遮挡更鲁棒（系统鲁棒性），计算更高效（系统高效性）的思路展开研究。首先，为了提高视频特征的完备性，本文先后提出功能互补的时空交互建模和时序互补建模方法。时空交互建模的目的是提高单帧特征的判别性；时序互补建模的目的是提高多帧特征融合后的视频特征的完整性；其次，为了提高视频特征对遮挡的鲁棒性，本文递进提出图像和特征层面的遮挡补全方法；最后，为了提高行人重识别的计算效率，本文设计了一个高效时空特征表示框架，在维持精度的同时大幅降低了计算开销。本文的主要研究内容及贡献如下：

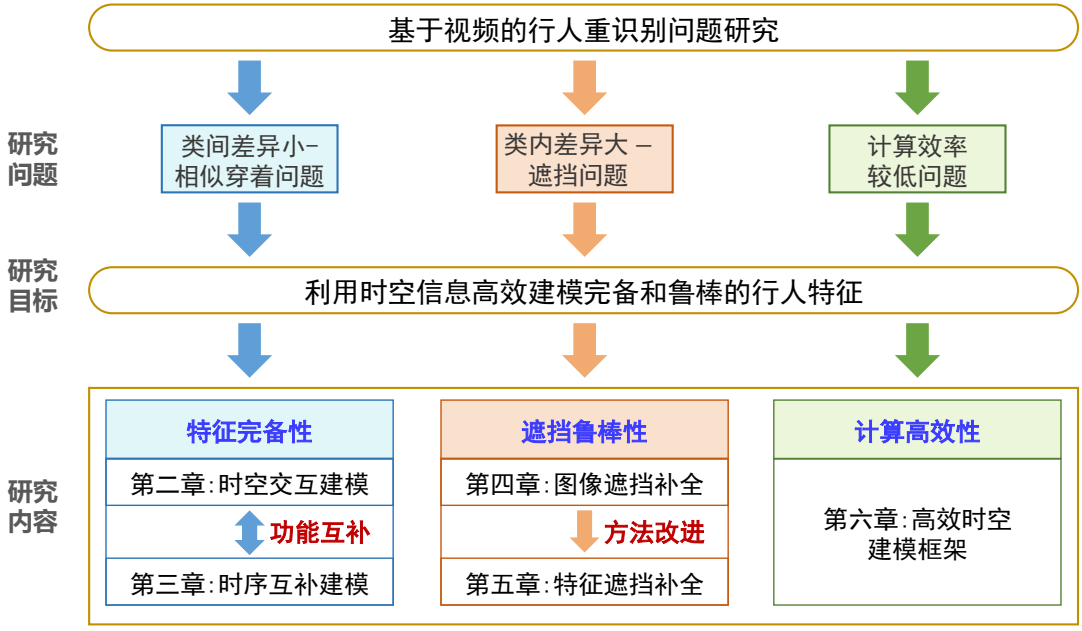


图 1.9 本文各章节之间的关系

- 提出时空交互的单帧判别特征学习方法

针对视频行人重识别方法的特征不完备问题，本文提出一种基于时空交互的特征学习方法。卷积操作在空间和时间上只处理一个局部区域，导致提取的特征缺乏全局信息，难以区分局部相似的不同行人。针对这个问题，该工作设计了一个全局建模的时空交互整合网络。具体地，首先学习一个部件检测单元分别提取每帧行人的头部、上身和下身等部件特征，然后分别对帧内不同部件特征进行空间交互，以及帧间相同部件特征进行时序交互，从而对每帧特征引入全局信息，提高了单帧特征的判别性。该工作的主要贡献是，提出一个新颖的时空特征交互机制，有效提高了单帧特征的判别性，进而增强多帧特征融合后的视频特征表示。

- 提出时序互补的多帧完整特征学习方法

上一个工作以相同操作处理每一帧图像，造成连续帧特征的高度相似和冗余。针对这个问题，本文提出一种时序互补建模方法，其核心思想是通过显式擦除其他帧已经关注的区域，引导当前帧的学习器挖掘新的互补区域，从而形成一个更完整的视频特征。该工作的主要贡献是，提出了时序互补建模的思想和方法，时序互补建模是得到视频行人完整表示的一个新颖手段。

上述两个工作可以有效结合。一方面，使用时空交互建模方法提高单帧特征的判别性；另一方面，使用时序互补建模方法引导连续帧关注不同的身体区域，以覆盖更完整的行人身体区域。综上分析，结合上述两个工作可以得到一个更完备的行人视频特征。

- 提出面向遮挡场景的时空补全学习方法

针对视频行人重识别的遮挡问题，本文提出一种时空遮挡补全方法。首先提出一个**图像**层面的时空补全方法，以渐进方式由粗到细逐步预测遮挡区域的像素。时空补全网络包括一个空间生成器和一个时序生成器。其中空间生成器利用单帧行人的身体结构信息初步复原遮挡区域的内容，时序生成器利用时序信息对初步复原的结果进行精细化。为了避免昂贵的图像生成，本文进一步提出一个**特征**层面的时空补全方法，直接修复遮挡区域的特征表示。该工作的主要贡献是，提出了一个针对遮挡问题的时空补全机制，有效提升了视频特征对遮挡的鲁棒性，大幅提升遮挡场景下的行人重识别精度。

- 提出高效时空建模框架

针对视频行人重识别方法的效率较低问题，本文提出一个基于双分支的高效时空建模框架。其中，细节分支处理原始分辨率的帧，用于保留细节的行人视觉特征；上下文分支处理降采样后的帧，用于捕捉行人长期的上下文特征。一方面，双分支框架建模了行人的不同尺度信息，具有较强的特征表示能力。另一方面，通过降低输入帧的分辨率，双分支框架大幅降低了视频处理所需计算开销，更利于部署于要求实时的场景。该工作的主要贡献是，提出了一个高效时空建模框架，达到了模型精度和速度的兼容。

1.7　本文组织结构

根据本文研究内容和方法，论文总共分为七个章节。章节安排如下：

第1章绪论。介绍了行人重识别任务的研究背景和挑战、国内外研究现状、以及公开的数据集与评价指标。最后给出了本文的研究内容和主要贡献。

第2章时空交互的单帧判别特征学习方法。从特征完备性的角度出发，解决传统卷积操作受限于局部建模的问题。提出了一个时空交互整合网络，并与已有先进工作进行了全面的对比和讨论。然而，交互整合网络无差别处理每一个视频帧，造成连续帧的特征冗余。所以下一章研究了如何减小同一个视频中连续帧的特征冗余，以提高视频特征的完整性。

第3章时序互补的多帧完整特征学习方法。针对上一章方法的特征冗余问题，提出了一个时序互补学习网络。该方法通过约束连续帧的特征互补，获得了一个更完整的视频特征。通过实验对比，验证了时序互补建模思想的有效性。然而，该方法主要针对非严重遮挡场景下的视频，没有考虑遮挡物的干扰和遮挡带来的时空信息损失。在下一章，本文研究了如何处理遮挡场景下的行人视频，以提取遮挡鲁棒的行人视频特征。

第4章面向遮挡场景的图像时空补全方法。由于遮挡物的干扰以及遮挡到来的时空信息损失，非遮挡行人重识别模型难以在遮挡场景有效工作。针对这个问题，提出了一个面向遮挡场景的图像时空补全网络，并从实验上验证了时空补全在遮挡场景的有效性。然而，图像补全的主要目标是生成视觉真实的图像，存在与行人重识别的目标不一致问题。在下一章对这一问题进行了研究。

第5章面向遮挡场景的特征时空补全方法。针对上一章方法的图像补全与行

人重识别目标不一致问题，提出了一个特征时空补全方法。该方法通过直接修复遮挡区域的特征表示，能够嵌入到行人重识别网络中，提高特征对遮挡的鲁棒性。通过在多个数据集与其他遮挡/非遮挡行人重识别方法进行对比，验证了特征补全相对于图像补全的优越性。然而上述方法都通过引入模块和更多的参数提高模型的准确性和鲁棒性，而未考虑行人重识别的效率问题。在下一章，本文研究了如何兼容重识别模型的精度和效率。

第6章高效时空建模框架。大部分行人重识别方法都关注于模型的精度，而忽略了模型的效率。针对这个问题，提出了一个双分支互补网络，以及设计了一个轻量级的时序核选择模块，在提升精度的同时显著提高了计算效率，并与已有的工作进行了详细的对比和讨论。

第7章总结与展望。对本文的研究工作做出总结，并对未来的研究方向和发展趋势做出展望。

第2章　时空交互的单帧判别特征学习方法

本章聚焦在视频行人重识别的完备性特征建模。针对卷积操作受限于局部建模的问题，本章设计了一个时空交互整合方法。该方法通过在每帧特征中引入全局信息，提高了单帧特征的判别性，进而增强多帧特征融合后的视频特征表示。

2.1　引言

再过去的几十年中，行人重识别技术凭借着它在智能视频监控等方面的巨大应用潜力受到了越来越多的关注。然而，由于人体的结构性以及监控场景中大量行人可能穿着相同款式衣服，不同行人的表观可能非常相似，从而使得行人重识别系统面临巨大挑战。随着深度学习的发展，深度学习技术在行人重识别任务上得到了广泛应用。近年来，基于深度学习的视频行人重识别方法取得了较好的性能，逐渐成为主流方法。

本章侧重于研究视频行人重识别中行人深度特征表达的关键技术。大部分方法采用一个“卷积神经网络+时序建模”框架：给定输入的行人视频，首先利用卷积神经网络提取每一帧的特征，然后通过池化操作 [43]、时序注意力操作 [44, 62] 或者循环网络 [48, 58, 59] 等融合所有帧的特征生成一个视频级别的行人特征表示。得益于深度神经网络 [85, 86] 强大的表达能力，这些方法取得了较好的性能。

然而，基于卷积神经网络的方法仍然具有以下局限性：卷积操作在空间和时间上只处理一个局部邻域，缺乏全局时空建模能力，导致提取的特征判别力不足。具体来说，由于卷积操作的时空局部性，局部区域的特征难以捕捉到全局行人身体信息。此外，由于训练样本有限，当卷积网络已经关注到一个可以在训练集中识别出目标行人的局部区域时，便不会再去挖掘其他的身体区域 [87–89]。因此，行人的最终卷积特征通常只关注一个最具表现力的局部区域，难以区分出局部相似的不同行人。因此，全局时空建模对于视频行人重识别任务是至关重要的。

为了实现行人序列的全局时空建模，本章提出了一个时空交互整合（Spatial-

Temporal Interaction-Aggregation, STIA）方法。在时空交互阶段，STIA生成一个时空关系图，旨在捕捉身体部件之间的空间和时序交互模式。其中空间交互模式建模了帧内不同身体部件之间的依赖关系，时序交互模式建模了帧间相同身体部件之间的依赖关系。通过这种方式，时空关系图融入了视频的全局时空信息；在时空整合阶段，STIA基于时空关系图整合视频序列的所有部件特征，得到一个全局时空特征表示。最后，将这个全局时空特征整合到每个视频帧中，提高了每帧特征的判别性。本章进一步提出了一个通道交互整合（Channel Interaction-Aggregation, CIA）方法。CIA通过建模视频特征图所有通道之间的语义交互，进一步增强每帧的特征表示。

基于提出的交互整合方法，本章分别设计了一个时空交互整合模块和通道交互整合模块。这两个模块都是轻量级的，只增加了不到10%的计算开销。这两个模块顺序组合成一个交互整合模块，可以嵌入到已有行人重识别方法中，使其具备全局建模能力。

本章接下来安排如下：章节2.2详细介绍了用于全局建模的交互整合特征学习方法；章节2.3给出了本章方法在公开数据集上的有效性分析；章节2.4将对本章内容进行小结。

2.2 全局建模的交互整合特征学习

在本节中，将对用于全局建模的交互整合特征学习方法进行详细介绍，包括对交互整合模块以及用于视频行人重识别的交互整合网络的介绍。

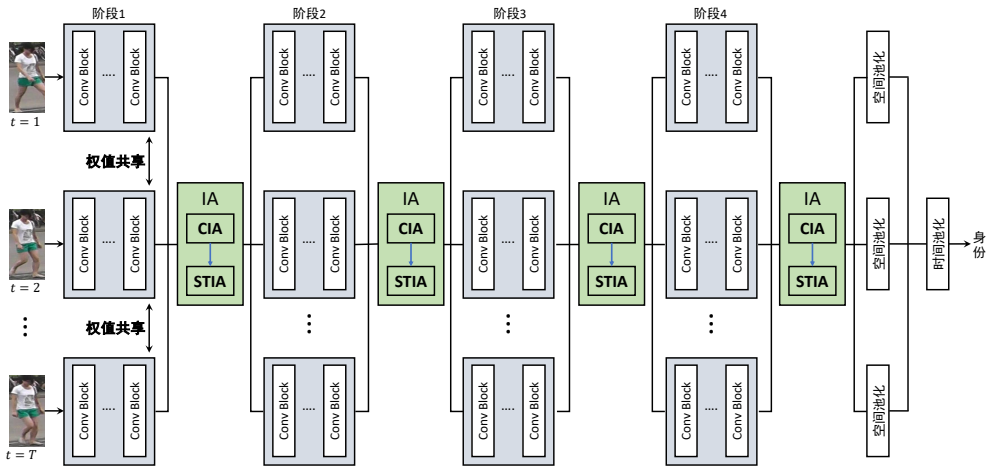


图 2.1 交互整合网络（Interaction-Aggregation Network, IANet）结构图

如图2.1所示，交互整合模块IA包含一个时空交互整合模块STIA，以及一个通道交互整合模块CIA。其中，STIA在时空维度上进行全局建模，CIA在通道维度上进行全局建模。由于STIA专注于“时空关系”和“不同身体部件之间”的建模，而CIA专注于“通道”和“相同部件”的建模。CIA可以通过建模不同通道之间的关系增强每个部件的特征，从而更有利于STIA学习不同部件之间的语义关系。因此，本章顺序组合CIA和STIA形成IA模块。如图2.1所示，IA模块能够嵌入到主干网络的任意深度中，形成用于行人重识别的交互整合网络。下面分别对时空交互整合模块，通道交互整合模块，以及交互整合网络予以介绍。

2.2.1　时空交互整合模块

行人序列的全局时空信息对于视频行人重识别任务至关重要。然而，主流方法受限于卷积操作的局部性，缺乏全局时空建模能力，导致提取的特征判别力不足。针对这个问题，本章设计了一个时空交互整合模块STIA，旨在生成一个具有全局时空信息的结构化特征表示。

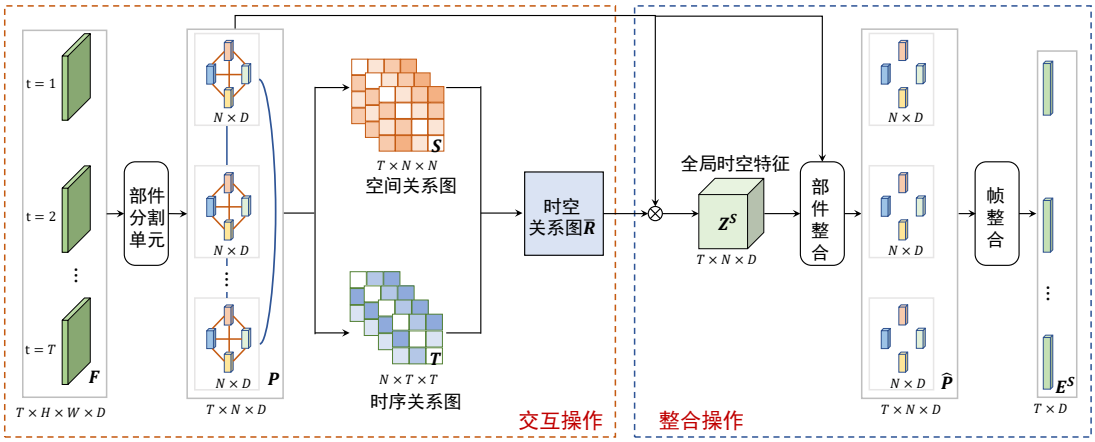


图 2.2 时空交互整合模块（Spatial-Temporal Interaction-Aggregation, STIA）结构图

如图2.2所示，给定一个视频卷积特征图$F \in \mathbb{R}^{T\times H\times W\times D}$，其中$T$、$H$、$W$和$D$分别表示特征图的帧数，高度，宽度和通道个数。STIA首先使用一个部件分割单元提取每帧的部件特征，其中每个部件特征对应一个指定的身体区域；然后将部件特征顺序经过交互操作和整合操作。其中，交互操作显式地建模部件之间的时空依赖关系，生成一个时空关系图；整合操作基于时空关系图整合所有关联部件的特征，获得一个蕴含全局时空信息的结构化特征，用以提高每帧特征的判别性。

2.2.1.1 交互操作

如图2.2所示，部件分割单元以视频特征图F作为输入，生成一个视频部件特征图$P \in \mathbb{R}^{T\times N\times D}$，其中$N$表示每帧划分的部件个数。部件分割单元的细节将在章节2.2.1.3介绍。为了建模行人序列的全局时空信息，交互操作需要同时捕捉$T \times N$个身体部件之间的空间和时序依赖关系。具体实现上，首先将行人序列的部件特征形式化为$P = \{p_{ij}|i = 1,\ldots,T, j = 1,\ldots,N\}$，其中$p_{ij} \in \mathbb{R}^D$表示第$i$帧的第$j$个部件特征。STIA模块考察两种类型的全局信息：空间关系和时序关系。

空间关系对帧内不同部件之间的语义交互进行建模。具体来说，STIA使用一个子网络学习每一帧内部任意两个部件之间的关联值，生成一个空间关系图$S = \{S_i\}_{i=1}^{T}$。其中$S_i \in \mathbb{R}^{N\times N}$表示第$i$帧的空间关系图，定义如下：

$$(S_i)_{jk} = W_r^T\left(\left[|p_{ij} - p_{ik}|, u\right]\right), \quad \text{其中 } u = GAP(F), k \neq j, \tag{2.1}$$

其中$|\cdot|$表示取绝对值操作，$[\cdot,\cdot]$表示向量连接操作，W_r为子网络的可学习权重。GAP表示全局平均池化操作，将视频特征图F沿着空间和时间维度执行平均池化，得到一个视频的粗糙全局特征u。使用这个粗糙的全局特征u，STIA能够在全局视角下估计身体部件之间的局部关系。

时序关系对帧间相同部件之间的语义交互进行建模。具体来说，STIA使用相同的子网络学习不同帧中相同部件之间的关联值，生成一个时序关系图$T = \{T_i\}_{i=1}^{N}$。其中$T_i \in \mathbb{R}^{T\times T}$为第$i$个部件的时序关系图，定义如下：

$$(T_i)_{jk} = W_r^T\left(\left[|p_{ji} - p_{ki}|, u\right]\right), \quad \text{其中 } k \neq j. \tag{2.2}$$

最终，STIA模块整合S和T生成一个时空关系图$R \in \mathbb{R}^{TN\times TN}$:

$$R_{ij} = \begin{cases} \left(S_{t_1}\right)_{n_1 n_2} & t_1 = t_2 \\ \left(T_{n_1}\right)_{t_1 t_2} & t_1 \neq t_2 \text{ and } n_1 = n_2 \\ 0 & t_1 \neq t_2 \text{ and } n_1 \neq n_2 \end{cases}, \tag{2.3}$$

其中$t_1 = i/N + 1$，$n_1 = i\%N + 1$，$t_2 = j/N + 1$，$n_2 = j\%N + 1$，R_{ij}表示第t_1帧的第n_1个部件与第t_2帧的第n_2个部件之间的关联值。最后，STIA使用Softmax函数对时空关系图进行归一化：

$$\overline{R}_{ij} = \begin{cases} \frac{\exp\left(R_{ij}\right)}{\sum_{k,R_{ik}\neq 0}\exp(R_{ik})} & R_{ij} \neq 0 \\ 0 & R_{ij} = 0. \end{cases} \tag{2.4}$$

STIA通过分解空间关系和时序关系，每个部件只与所有$N \times T$个部件中的$N+T-1$个部件关联，大幅度降低了交互操作的计算量。

2.2.1.2　整合操作

整合操作根据时空关系图$\overline{R}$整合视频的所有部件特征。实现上，首先将P调整为$TN \times D$大小，然后执行$\overline{R}$和P的矩阵相乘操作，得到全局时空特征Z^S:

$$Z^S = \overline{R}P. \tag{2.5}$$

Z^S被调整为$T \times N \times D$大小，以与输入视频部件特征图的大小保持一致。从等式2.5可以看出，Z^S的每个部件特征都融入了视频的全局时空信息，具有更强的判别力。

然后，STIA设计了一个部件整合单元，将全局时空特征与初始部件特征进行整合，获得一个更新的部件特征$\hat{P} \in \mathbb{R}^{T \times N \times D}$：

$$\hat{p}_{ij} = W_{pu}^T \left(\left[p_{ij}, Z_{ij}^S \right] \right), \tag{2.6}$$

其中$W_{pu} \in \mathbb{R}^{2D \times D}$为将连接向量投影到原始特征空间的可学习权重。

最后，STIA设计了一个帧整合单元，对$\hat{P}$执行全局空间池化操作，并整合粗糙的视频全局特征u，获得一个具有全局时空信息的结构化特征$E^S \in \mathbb{R}^{T \times D}$:

$$E_i^S = W_{fu}^T \left(\left[\frac{\sum_j \hat{p}_{ij}}{N}, u \right] \right), \tag{2.7}$$

其中$W_{fu} \in \mathbb{R}^{2D \times D}$为将连接向量投影到原始特征空间的可学习权重。

2.2.1.3　部件分割单元

本节详细介绍STIA的部件分割单元。部件分割单元的目的是定位到行人的身体部件区域。目前方法 [10, 24, 90] 通常借助一个外部的部件检测网络。这类方法会导致行人重识别系统过于耗时，不利于真实场景的部署。不同于已有方法 [10, 24, 90]，本章设计了一个轻量的空间注意力子网络去定位行人部件。具体来说，给定输入的视频特征F，首先使用一个卷积层生成多个身体部件的空间注意力图$A \in \mathbb{R}^{T \times H \times W \times N}$：

$$A = \sigma \left(W_a * F + b_a \right), \tag{2.8}$$

其中σ表示Sigmoid激活函数，$*$表示卷积操作，W_a和b_a分别表示卷积核的权重和偏差。之后将生成的注意力图作用到视频特征上，得到视频部件特征P：

$$p_{ij} = \frac{1}{HW}\left(\sum_{hw} A_{ihwj} F_{ihw}\right). \tag{2.9}$$

但是，在缺乏指导的情况下，部件分割单元生成的注意力图可能会关注到背景区域。为此，本方法使用行人身体部件掩码M指导注意力图A的生成。实现上，使用开源的的行人分割模型提取每个输入帧的部件掩码M，然后将M调整到与注意力图A相同的尺度，最后最小化部件分割损失。部件分割损失定义为M和A的二值交叉熵，形式化如下：

$$L_p = -\left[M \log A + (1 - M) \log (1 - A)\right]. \tag{2.10}$$

2.2.2 通道交互整合模块

现有行人重识别方法通常堆叠多个卷积层提取行人特征。随着层数的增加，小尺度的身体部件（例如鞋子）很容易在特征图中褪去。但是这些小尺度部件十分有助于区分具有微小差异的不同行人。文献 [91] 指出大多数高层特征图的通道都响应一个特定的部件。基于这个观点，本章构建了一个通道交互整合模块CIA，旨在整合所有通道中语义相似的特征。通过整合来自其它通道的特定部件信息，CIA模块能够增强每个部件的特征表示。

2.2.2.1 交互操作

CIA的结构如图2.3所示。在交互阶段，CIA以视频特征图F作为输入，显式建模F不同通道之间的语义关系，生成一个通道关系图。实现上，首先将F调整为$TD \times HW$大小，然后对F和F的转置执行矩阵乘法，并利用Softmax函数归一化结果得到通道关系图$C \in \mathbb{R}^{TD \times TD}$。形式化地，任意两个通道之间的语义相似值的计算如下：

$$C_{ij} = \frac{\exp\left(f_i^T f_j\right)}{\sum_{k=1}^{TD} \exp\left(f_i^T f_k\right)}, \tag{2.11}$$

其中$f_i, f_j \in \mathbb{R}^{HW}$分别表示$F$的第$i$个和第$j$个通道。

2.2.2.2 整合操作

整合操作基于通道关系图整合不同通道的特征表示。具体来说，首先执行通道关系图C和视频特征图F的矩阵相乘操作，获得特征图$Z^C \in \mathbb{R}^{TD \times HW}$:

$$Z^C = CF. \tag{2.12}$$

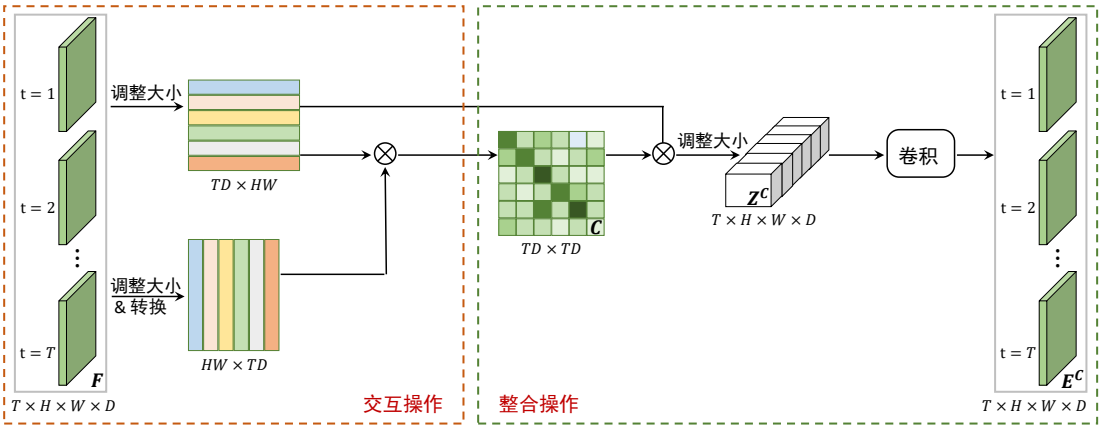


图 2.3 通道交互整合模块（Channel Interaction-Aggregation, CIA）结构图

Z^C被调整为$T \times H \times W \times D$大小，以与输入视频特征图的大小保持一致。最后，将$Z^C$通过一个卷积层获得最终的特征图$E^C$。

CIA根据通道关系图整合不同通道之间的语义，STIA根据时空关系图整合不同部件之间的语义，因此这两个模块是功能互补的。与STIA类似，CIA可以自适应调整输入的视频特征图，提高特征的判别性。

2.2.3　用于行人重识别的交互整合网络

用于视频行人重识别的交互整合网络IANet的结构如图2.1所示。IANet使用ResNet50 [85] 作为主干网络，使用时序平均池化（Temporal Averae Pooling）融合多帧特征得到一个视频特征表示。考虑到计算复杂度，本章只将IA模块嵌入到特征图发生下采样的瓶颈层。不同于已有工作 [48, 58, 59] 只在网络顶层构建时序依赖关系，IA模块能够放置在底层阶段捕捉更丰富的全局时空依赖。当输入序列的帧数为1时，IANet也可以用于图像行人重识别任务。

2.2.3.1　损失函数

训练交互整合网络的目标函数包括：行人身份监督的交叉熵损失L_{ce}，困难样本三元组损失L_{tri}，以及指导部件分割单元的损失L_p （等式2.10）。

交叉熵损失L_{ce}　交叉熵损失通过行人重识别网络输出的类别概率与真实身份标签进行损失评估。具体来说，假定训练集有K个行人身份，将一个序列输入行人重识别网络提取其特征f，然后经过一个全连接（Fully Connect，FC）层输出该序列的身份预测向量$z = [z_1, z_2, \ldots, z_K]$。最后利用Softmax计算输入序列属

于第k个身份的概率：$p(k)=\frac{\exp(z_k)}{\sum_{i=1}^{K}\exp(z_i)}$。最终交叉熵损失定义为：

$$L_{ce}=-\sum_{k=1}^{K}q(k)\log p(k), \tag{2.13}$$

其中q表示输入序列的真实身份标签，如果输入序列的身份标签为y，则$q(y)=1$，$q(k|k\neq y)=0$。

困难样本三元组损失L_{tri} 三元组损失的目标是在一定距离上分开正负样本。然而大多数三元组为简单样本，不利于模型的训练。本文使用文献 [20] 提出的困难样本三元组损失，即对于每个目标样本，在一个训练批次（Batch）内挑选出距离最远的正样本对和距离最近的负样本对来训练网络，定义如下：

$$L_{tri}=\sum_{i=1}^{C}\sum_{a=1}^{L}\left[\alpha+\max_{p=1,\ldots,L}d\left(f_a^i,f_p^i\right)-\min_{\substack{j=1,\ldots,C\\ n=1,\ldots,L\\ j\neq i}}d\left(f_a^i,f_n^j\right)\right]_+, \tag{2.14}$$

其中，C表示Batch内行人身份个数，L表示Batch内每个行人的序列个数，f_j^i表示Batch内第i个行人的第j个序列，d表示余弦距离，超参数α为距离阈值。

整体目标函数定义为：

$$L_{all}=(L_{ce}+L_{tri})+\lambda_1 L_p. \tag{2.15}$$

λ_1是用来平衡各个损失的超参数。

2.2.4 与其他相似模块的讨论

本节简要讨论了IA模块与另外两个相似模块 Nonlocal [92] 和 Sequeeze-and-Excitation [93] 之间的关系。实验比较见章节2.3.3。

与Nonlocal比较 IA和Nonlocal都可以视作图网络模块。与Nonlocal相比，IA主要有以下优势：（1）Nonlocal可以视为输入特征图所有空间位置特征的一个稠密连接图。它需要计算一个稠密的关系矩阵，对于大尺度的特征图非常不友好。相比之下，STIA对于任意大小的输入特征图都有N个图节点，具有更小的计算复杂度。（2）Nonlocal只能捕捉固定位置之间的外观相似关系。STIA通过对远距离的身体部件之间执行更高阶的关系建模，能够增强局部区域特征的判别性，提高模型对相似行人的区分能力。（3）NonLocal只能建模空间特征之间的语义关系。而本章的CIA能够捕捉到通道特征之间的语义关系。CIA与STIA功能互补，有利于增强微小但重要的身体部件信息。

与Sequeeze-and-Excitation的比较　CIA与Sequeeze-and-Excitation（SE）都旨在建模通道之间的依赖关系。然而，SE计算一个通道注意力图，强调更具表现力的通道特征，可能会忽略细小但重要的部件。相比之下，CIA整合了所有通道中语义相似的特征，能够增强所有部件的特征表示。

2.3　实验分析

本节分别在MARS [43] 、DukeMTMC-VideoReID [76] 和LS-VID [94] 三个视频数据集上对本章提出的交互整合方法进行评测。为了进一步验证本章方法的泛化性，本节还在三个图像数据集Market-1501 [29]，DukeMTMC [74] 和MSMT17 [41] 上进行测试。本节接下来分别介绍了实验细节、消融实验、与当前先进方法的对比实验以及可视化分析。

2.3.1　实现细节

视频行人重识别上的实现细节　本章采用Adam优化器 [95] 进行网络优化，共训练150个回合。初始学习率设为3×10^{-4}，每训练40个回合后下降10倍。输入帧的大小设为256×128。训练使用的批量大小为32，其中包含8个行人类别，每个类别包含4个视频片段。在训练时，输入的视频片段的长度应该相同，并且每一帧应该以相同的概率被采样。因此，对于每个原始视频，以8帧为间隔随机采样4帧形成一个视频片段。本章使用水平翻转和随机裁剪作为数据增广。超参数λ_1设置为0.5，三元组损失的距离阈值设置为0.3。

图像行人重识别上的实现细节　对于图像行人重识别任务，本章方法共训练60个回合，批量大小为64，初始学习率设为3.5×10^{-4}，每训练20个回合后下降10倍。其他的训练设置与视频行人重识别任务一致。

2.3.2　消融实验

为了验证IANet各个模块的有效性，本节分别在视频数据集和图像数据集上采取一系列消融实验。首先引入一个常用的基准模型（Baseline），使用ResNet-50作为特征提取器，通过交叉熵损失和三元组损失联合训练。为了公平比较，Baseline使用与RFCnet相同的训练细节。实验结果见表2.2和表2.1。

STIA模块的有效性　如表2.1和2.2所示，STIA模块在视频和图像行人重识别任务上带来了一致的性能提升。在图像行人重识别上，STIA模块带来了2%左

表 2.1 本章方法在视频行人重识别上MARS和DukeMTMC-Video数据集上的消融实验

方法	MARS		DukeMTMC-Video	
	mAP (%)	top-1 (%)	mAP (%)	top-1 (%)
Baseline	83.5	88.2	94.5	95.0
STIA-spatial	84.6	88.9	95.3	96.0
STIA-temporal	84.6	89.1	95.0	95.8
STIA	**84.9**	**89.6**	**95.7**	**96.2**
CIA	84.5	89.1	95.6	96.0
IA	**85.0**	**90.2**	**96.1**	**96.9**
IA-wo-L_p ($stage_2$)	84.5	88.2	95.3	96.2
IA ($stage_2$)	**85.0**	**90.2**	**96.1**	**96.9**
IA ($stage_3$)	84.5	89.1	96.0	96.8
IA ($stage_{23}$)	**85.3**	90.0	95.9	96.3

右的mAP增益。另外，本节研究了STIA在空间、时间和时空维度上对视频行人重识别性能的影响。在表2.1中，STIA-spatial只对帧内的空间关系进行建模，STIA-temporal只对帧间的时序关系进行建模。可以观察到，STIA-spatial和STIA-temporal都优于Baseline的性能，并且空间和时序关系集成的STIA模块能够进一步提升重识别性能，表明了全局空间建模和全局时序建模的有效性和互补性。

CIA模块的有效性 本节进一步评测了CIA模块的有效性。如表2.1和2.2所示，CIA模块在视频和图像数据集上一致提升了2%的mAP和1%的top-1。实验表明了整合相似的通道特征能够有效增强特征表示能力。如表2.1和2.2所示，IA模块集成了STIA和CIA，进一步提升了行人重识别性能，说明了STIA和CIA模块的互补性。

STIA和CIA模块的组合方式 本节比较了三种STIA和CIA模块的组合方式：顺序组合STIA和CIA（STIA-CIA），顺序组合CIA和STIA（CIA-STIA），以及两个模块的并行使用（STIA+CIA）。实验结果如表2.3所示，CIA-STIA在视频和图像重识别任务上都取得了最好的性能。这主要是由于STIA和CIA计算互补的语义关系，其中STIA专注于“时空”和“不同部件”的建模，而CIA专注于“通道”和“相同部件”的建模。CIA通过建模通道关系增强了每个部件的特征，更利于STIA学习不同部件之间的语义关系。因此，本章顺序组合CIA和STIA形

表 2.2 本章方法在图像行人重识别上Market-1501和DukeMTMC数据集上的消融实验

方法	Market-1501		DukeMTMC	
	mAP (%)	top-1 (%)	mAP (%)	top-1 (%)
Baseline	84.5	93.4	76.2	87.8
STIA	86.5	94.6	78.2	89.1
CIA	86.3	94.4	78.1	88.9
IA	**86.7**	**94.8**	**78.9**	**89.2**
IA ($stage_1$)	85.6	93.8	77.3	88.2
IA ($stage_2$)	86.7	**94.8**	78.9	89.2
IA ($stage_3$)	**86.8**	94.3	**79.3**	**89.3**
IA ($stage_4$)	85.1	93.8	76.7	87.6
IA-wo-L_p ($stage_{23}$)	87.3	94.7	78.5	88.9
IA ($stage_{23}$)	**88.2**	**95.0**	**79.5**	**89.6**

表 2.3 STIA和CIA不同组合方式的性能对比

方法	MARS		Market-1501	
	mAP (%)	top-1 (%)	mAP (%)	top-1 (%)
顺序组合STIA和CIA（STIA-CIA）	84.5	88.2	85.6	94.3
顺序组合CIA和STIA（CIA-STIA）	84.4	88.6	85.9	94.4
并行组合CIA和STIA （STIA+CIA）	**85.0**	**90.2**	**86.7**	**94.8**

成IA模块。

IA模块的放置位置　表2.2比较了放置在ResNet50不同卷积阶段的IA模块，其中IA模块被嵌入到卷积阶段的最后一个残差块之前。如表2.2所示，IA模块能够一致提升基准模型的性能，并且放置在第二个卷积阶段和第三个卷积阶段更为有效。一个可能的解释是，第一个阶段的卷积特征图表现力较弱，不足以提供精确的语义线索；而第四个阶段的卷积特征过于抽象，难以整合语义相关的时空/通道特征。因此，本章只考虑将IA模块放置到第二个和第三个卷积阶段。表2.2和2.1展示了放置更多IA模块的结果。可以观察到，对于图像行人重识别，更多的IA模块持续提升重识别性能。对于视频行人重识别，添加更多的IA模块并不会带来显著的性能增益，因此本方法只在主干网络的第二个卷积阶段嵌入一个IA模块用于视频行人重识别。

表 2.4 IA模块与不同主干网络结合的性能

主干网络	方法	MARS		Market-1501	
		mAP (%)	top-1 (%)	mAP (%)	top-1 (%)
ResNet18	Baseline	74.9	83.7	72.0	89.1
	ResNet18+IA模块	**76.4**	**84.8**	**74.9**	**89.8**
ResNet34	Baseline	78.4	85.9	74.8	89.5
	ResNet34+IA模块	**79.7**	**86.9**	**78.5**	**91.0**
Inception	Baseline	72.7	82.2	72.2	88.1
	Inception+IA模块	**74.1**	**83.6**	**75.2**	**89.7**

不同主干网络下IA模块的有效性 本节首先测试了IA模块与ResNet18 [85]和ResNet34 [85] 组合的结果。如表2.4所示，IA模块能够取得一致的性能提升。本节进一步测试了IA模块在非残差网络Inception [96] 上的有效性，可以观察到与残差结构一致的现象。上述实验表明IA模块可以嵌入到多种卷积架构中一致提升行人重识别的准确性。

部件分割约束L_p的有效性 本节测试了部件分割约束L_p的有效性，引入了一个比较模型IANet-wo-L_p，该模型去除等式2.15中的部件分割损失L_p。如表2.1和2.2所示，去除L_p导致行人重识别精度下降，说明了部件分割约束的有效性。图2.4可视化IANet和IANet-wo-L_p学习到的部件分割结果。从图2.4中可以观察到，IANet-wo-L_p生成的多个注意力图是杂乱无章的，并且会集中到一个相同的区域。在这种情况下，IANet-wo-L_p无法建立不同身体部件之间的依赖关系，导致性能降低。

2.3.3 与相似方法的比较

本节展示了IA模块与章节2.2.4提到的Nonlocal（NL）和 Squeeze-and-Excitation（SE）模块的对比结果。实验结果如表2.5所示。

STIA与NL的对比 与NL方法相比，STIA模块以更少的计算开销取得了更高的精度。这主要是因为NL很难推理出不同部件之间的语义关系。相反，STIA模块通过显式关联远距离的不同部件，更容易捕捉到部件之间的长距离语义关联，从而通过全局时空建模提高局部区域特征的判别性。

CIA与SE的对比 如表2.5所示，CIA模块的性能显著优于SE模块。相比

表 2.5 IA模块与NonLocal（NL）和Squeeze-and-Excitation（SE）模块的性能对比，GFLOPs表示模型平均处理一帧所需的浮点运算次数，Param.表示模型的参数量

方法	MARS			
	Param.	GFLOPs	mAP (%)	top-1 (%)
ResNet50（Baseline）	23.5M	4.06	83.5	88.2
ResNet50+NL [92]	24.0M	4.86	84.0	88.8
ResNet50+STIA	24.0M	4.06	**84.9**	**89.6**
ResNet50+SE [93]	23.5M	4.06	83.5	88.7
ResNet50+CIA	23.7M	5.26	**84.5**	**89.1**
ResNet50+IA（IANet）	24.3M	5.27	**85.0**	**90.2**

于Baseline，SE模块只能带来很少的性能增益，而CIA模块在mAP和top-1上分别提升了1.0%和0.9%。实验结果表明相比常用的通道注意力机制，整合相似的通道特征的CIA更利于增强特征表示能力。但是，CIA的计算量略高于SE，增加的计算量主要来自SE的通道关联图的计算。由于这个操作可以通过矩阵乘法实现，SE在GPU上只占用很少的时间。

2.3.4 与当前先进方法的对比

本节分别在视频和图像行人重识别任务上将IANet与当前先进的方法进行比较，然后比较了IANet在视频和图像任务上的表现。

2.3.4.1 视频行人重识别上的方法对比

本节将IANet与当前先进的视频行人重识别方法进行比较。对比方法分为两组：基于集合建模的深度学习方法和基于时空建模的深度学习方法。实验结果如表2.6所示。从表2.6可以总结出：（1）相比于集合建模的方法 [45–47, 76]，本章方法能够取得更优的性能：在MARS上mAP大约提升了6.0%，说明了时序信息对于视频行人重识别任务的有效性。（2）最近工作 [49, 52] 使用三维卷积或者NonLocal建模视频的时序信息，但是这两类方法都有较高的计算复杂度。本章的IANet能够以更少的计算开销取得更优的性能：在MARS上mAP大约提升了5%，在DukeMTMC-VideoReID上mAP提升了2%。

LS-VID数据集存在大量着装非常相似的行人，具有严重的类间相似问题。从表2.6可以观察到，已有方法M3D[52]和 GLTP[94]在LS-VID数据集上的精度很

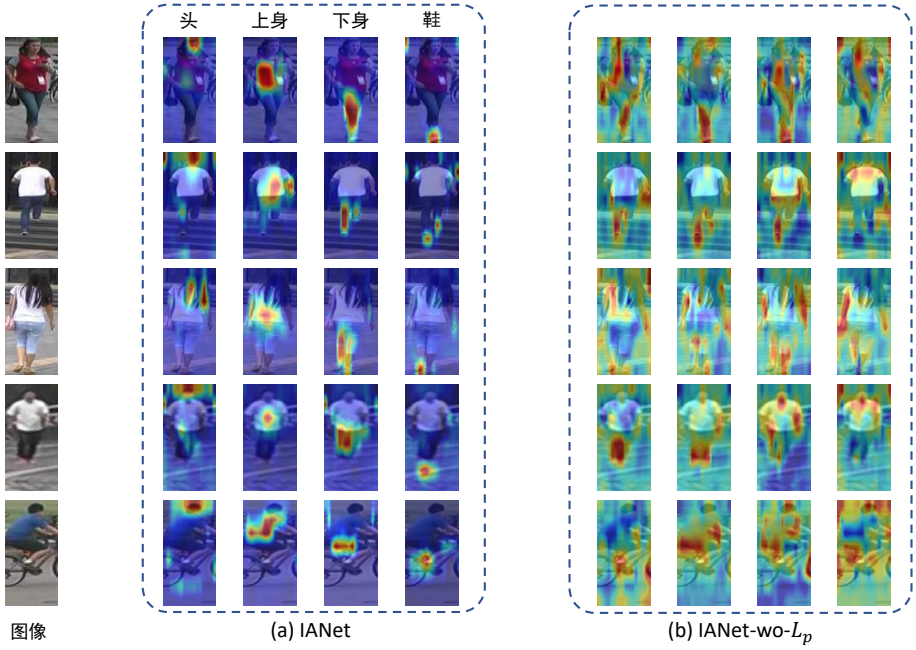


图 2.4 IANet和IANet-wo-L_p生成的部件分割注意力图

低，而IANet取得了远高于M3D和GLTP的性能，在mAP和top-1上提升了高达20%。性能的大幅提升归功于IA模块能够捕捉全局的时空信息，有利于区分LS-VID数据集中大量的相似不同行人。

2.3.4.2 图像行人重识别上的方法对比

表2.7总结了本章方法与当前先进的图像行人重识别方法的对比结果。对比的方法分为两组：基于全局特征的深度学习方法和基于局部特征的深度学习方法。如表2.7所示，IANet在Market-1501、DukeMTMC和MSMT17数据集上都能取得最优的结果。（1）IANet的性能显著优于基于全局特征的方法 [99–102]，在mAP上大约提升了6%。（2）一些基于局部特征的方法 [24, 103] 引入一个部件检测网络，导致行人重识别系统过于复杂和耗时。IANet以更小的计算开销取得了更高的性能。（3）另外一些基于局部特征的方法 [14, 23] 使用轻量级的注意力子网络 [104] 定位行人的身体部件。相比于这类方法，IANet能够取得6%的mAP提升，进一步验证了IANet的有效性。

MSMT17是LS-VID的同源数据集，在LS-VID的每个行人序列中随机选择一帧组成MSMT17。因此，MSMT17数据集也存在严重的类间相似问题。从表2.7可以观察到，IANet在MSMT17数据集上显著优于当前方法，在mAP和top-1上分别提升了25.9%和20.6%。实验结果表明了IANet在具有严重类间相似问题的场景

表 2.6 本章方法与现有视频行人重识别方法在MARS、DukeMTMC-Video和LS-VID数据集上的性能对比

	方法	MARS		Duke-Video		LS-VID	
		mAP (%)	top-1 (%)	mAP	top-1	mAP	top-1
集合建模	QAN [44]	51.7	73.7	–	–	–	–
	STAN [46]	65.8	82.3	–	–	–	–
	EUG [76]	67.4	80.8	78.3	83.6	–	–
	RQEN [45]	71.7	77.8	–	–	–	–
	TAFD [47]	78.2	87.0	–	–	–	–
时空建模	M3D [52]	74.1	84.4	–	–	40.1	57.7
	Snippet+OF [61]	76.1	86.3	–	–	–	–
	V3D [49]	77.0	84.3	–	–	–	–
	GLTP [94]	78.5	87.0	93.7	96.3	44.3	63.1
	CoSAM [97]	79.9	84.9	93.7	96.2	–	–
	IANet	**85.0**	**90.2**	**96.1**	**96.9**	**69.2**	**80.1**

下的优越性。

2.3.4.3　视频和图像的性能对比分析

从表2.6和表2.7可以观察到：IANet在LS-VID视频数据集的性能显著高于同源的MSMT17图像数据集，在mAP上提升了高达10%。这主要是因为图像所包含的信息非常有限，单帧图像的姿态、视角、质量等都会对行人重识别准确率产生较大的影响。如果一个摄相机视角下的行人信息出现了缺失，很难找到其他有效信息帮助识别。比如，正面的视角下无法捕捉到背包等细粒度信息，而这些细粒度线索非常有助于区分着装相似的不同行人。而视频行人重识别的研究对象是图像序列，每个行人在一个摄像机视角下包含一组图像，具有更多的可利用信息，包括行人的多种姿态和视角、运动等信息。因此，相比于图像，视频序列更有可能涵盖行人的全部信息，从而更有利于区分出具有微小差异的不同行人，在具有大量相似行人的场景下能够取得更好的性能水平。

2.3.5　可视化分析

为了进一步验证方法的有效性，本节分别对IANet学习到的空间注意力图、

表 2.7 本章方法与现有图像行人重识别方法在三个数据集上的性能对比

	方法	Market-1501		DukeMTMC		MSMT17	
		mAP (%)	top-1 (%)	mAP	top-1	mAP	top-1
全局特征	PDC [98]	63.4	84.1	–	–	29.7	58.0
	MLFN [99]	74.3	90.0	62.8	81.0	–	–
	KPM [100]	75.3	90.1	63.2	80.3	–	–
	Mancs [101]	82.3	93.1	71.8	84.9	–	–
	Group [102]	81.6	93.5	69.5	84.9	–	–
	GLAD [41]	–	–	–	–	34.0	61.4
局部特征	PSE [103]	69.0	87.7	62.0	79.8	–	–
	MGCAM [23]	74.3	83.7	–	–	–	–
	SPReID [24]	81.3	92.5	70.9	84.4	–	–
	RPP [14]	81.6	93.8	69.2	83.3	–	–
	CASN [89]	82.8	94.4	73.7	87.7	–	–
	IANet（本章方法）	**88.2**	95.0	**79.5**	**89.6**	**59.0**	**78.0**

空间和通道关系图进行可视化分析。

空间注意力图可视化 图2.4可视化了部件分割单元学习到的空间注意力图。可以观察到，不同的空间注意力图能够关注不同的局部身体部件：头部、上半身、下半身和鞋子。在各种具有挑战的情况下，比如小尺度（第二行）、运动模糊（第四行）以及剧烈的姿态变化（最后一行），空间注意力图都能够自适应定位到行人的身体部件。由于卷积操作不能直接建模远距离部件之间的依赖关系，局部特征一般缺乏全局语义，判别能力较弱。与之相反，IA模块根据注意力图可以生成指定的部件特征，并显式建模不同部件之间的全局时空关系，提高了每个局部部件特征的判别性。

STIA和CIA关系图可视化 本节分别可视化了STIA的空间关系图以及CIA的通道关系图。图2.5可视化了初始部件特征P，空间交互整合更新后的部件特征$\hat{P}$，以及空间关系图S。可以观察到，对于上衣相似的两个行人，初始上半身特征很难区分这两个行人。空间关系图S存储了全局空间信息。如图2.5所示，对于上半身部件，S分配给鞋子和头部较大的关联值。由于鞋子和头部具有较高的区分度，通过S更新的上半身特征能够区分这两个行人。此外，我们观察到头部和鞋趋向于在S中呈现较高的值。这是由于头部和鞋子通常比衣服更具区分

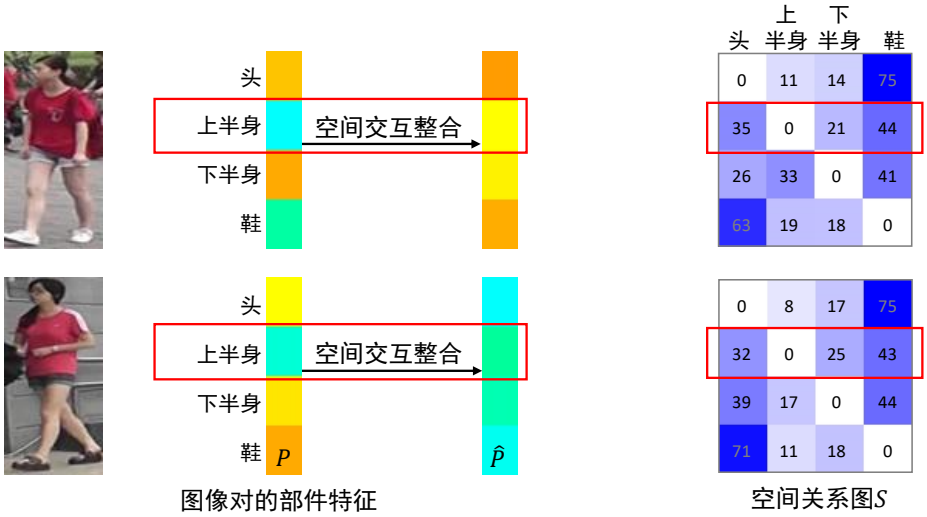


图 2.5 空间关系图可视化结果

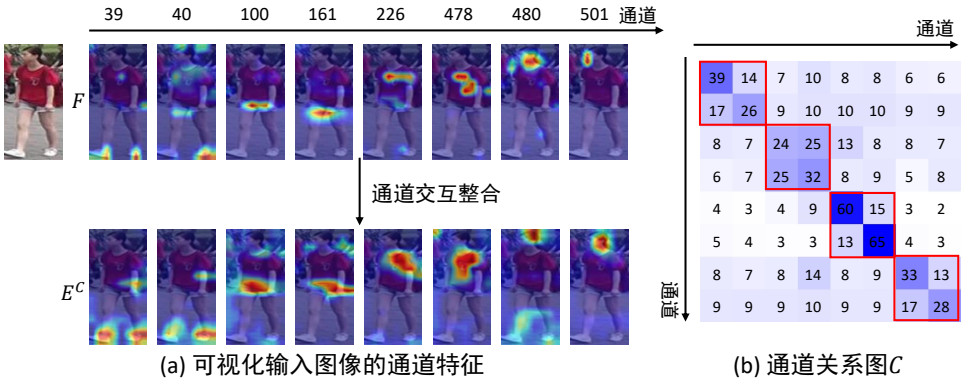


图 2.6 通道关系图可视化结果

度。因此，STIA学到对头部/鞋子与其他部件之间的关联分配较高的值，以便其它身体部件能够整合头部/鞋子的信息提高其判别性。

图2.6可视化了CIA的通道关系图C。为了清楚地可视化通道关系图，本节随机选择8个通道并可视化它们的初始特征F，CIA更新的特征E^C，以及这8个通道之间的关系图。如图2.6所示，关注同一个身体部件的通道特征趋向于具有更高的关联值。通过C在通道间传播特征之后，每个通道融入了来自其它通道的指定部件信息。如图2.6所示，经过CIA模块，通道特征能够关注到特定部件的更多区域，增强了对应部件的特征表示能力。

2.4　本章小结

面向视频行人重识别的完备性特征表示，本章提出了一种新颖的用于全局建模的交互整合模块。本章的工作表明，全局建模对行人重识别任务是有价值

的，而交互整合机制是实现全局建模的一个有效手段。得益于提出的交互整合方法，行人的局部部件特征能够融入全局信息，从而更具判别性，特别是降低了相似行人的区分难度。在多个数据集的实验表明交互整合模块能够在一定程度上解决类间相似问题，显著提高行人重识别的性能。

本章的IANet仍然具有一定的局限性。虽然IANet通过全局建模在一定程度上减小了相似区域的特征混淆，但是IANet以相同的操作处理每一帧，导致连续帧特征的高度冗余，并且趋向于关注同一个局部区域。这就造成了，IANet提取的视频特征依旧无法覆盖一个完整的行人身体区域，始终受限于区分具有微小差异的不同行人。在下一章工作中，将从时序互补建模的角度进一步提高视频特征的完备性。

第3章　时序互补的多帧完整特征学习方法

上一章提出的方法虽然提高了视频中每帧特征的判别性，但是仍然存在多帧特征冗余的问题。这些冗余的特征趋向于关注同一个显著但是局部的区域，导致最终的视频特征无法覆盖一个完整的行人身体区域。针对这个问题，本章提出了一种“时序互补建模”思想：通过引导连续帧关注互补的身体区域，以提高多帧整合的视频特征的完整性。

3.1　引言

一个完备的视频行人特征至少应该满足两个条件：（1）判别性：特征应该能够充分描述所关注区域的视觉信息。（2）完整性：特征应该能够覆盖一个完整的行人身体区域。然而现有方法集中在提高特征的判别性，而忽视了其完整性。当前先进的视频行人重识别方法主要采用“时序增强建模”的思想，即利用时序关系实现不同帧特征之间的相互增强。具体来说，已有方法采用循环网络[48, 58]、三维卷积[49, 51]以及包括上一章交互整合IANet方法的图网络[53, 66]，鼓励视频帧之间的信息传播。通过这种方式，每帧特征融合了其他帧中的关联信息，增强了其表征能力，因而更具判别力。但是这类方法存在一个内在缺陷，即无法建模一个完整的行人表示，这主要是由以下两个原因造成的。

首先，对于每一个视频帧，现有方法通常只关注一个显著但是局部的区域。事实上，当模型已经关注到一个可以识别目标行人的局部区域时，模型便不会再去关注其他区域[87, 88]。如图3.1所示，已有方法[25]几乎把所有的注意力都集中到黄色上衣上，忽视了短裤和鞋子。但是，短裤和鞋子也是非常重要的视觉线索，特别是用于区分其他黄色上衣的行人。尽管上一章的IANet通过全局建模在每帧特征中融入了全局身体信息，在一定程度上缓解了局部相似区域的特征混淆。但是，IANet所关注的局部区域依旧主导了行人特征表示。因此，IANet并没有充分利用全部的细粒度视觉线索，使得相似行人的区分依旧极具挑战性。

其次，现有方法以相同的操作处理每一帧图像，导致相邻帧的特征高度相似和冗余。这个问题在基于时序增强建模的工作（包括上一章的IANet）上尤为

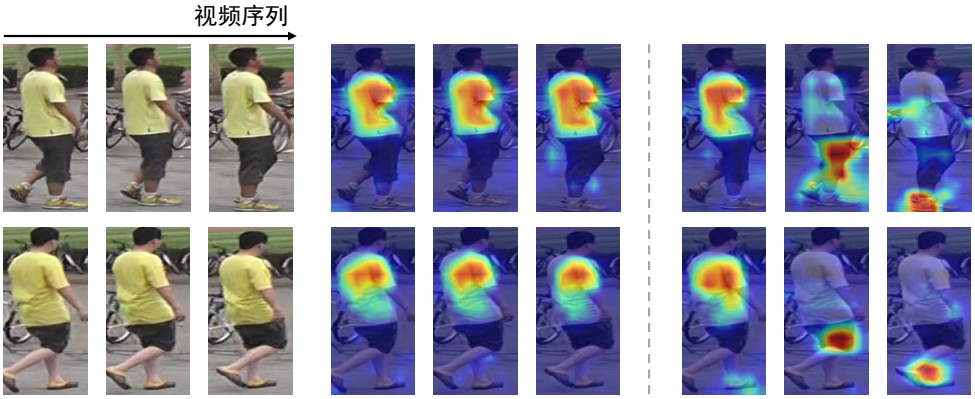


图 3.1 视频序列的特征激活图

严重。具体来说，这类方法利用时序信息在不同帧之间传播特征，因而每帧特征包含了来自其它帧的相互增强的信息。这种时序利用方式虽然提高了单帧特征的判别性，但同时造成了不同帧的特征更为相似。这些高度相似的特征会关注到同一个显著但是局部的区域，导致多帧整合后的视频特征仍难以区分局部相似的不同行人。如图3.1(b)所示，序列的连续三帧都关注到黄色上衣区域，最终的视频特征无法区分图中的这两个行人。综合上述分析，本章希望从另一个角度利用时序信息，鼓励对连续帧挖掘“互补”的视觉线索，从而形成一个更完整的行人视频特征表示。

为了实现上述目标，本章提出了“时序互补建模”思想，并设计了一个时序互补学习网络（Temporal Complementary Learning Network，TCLNet）。首先提出了一个时序显著性擦除（Temporal Saliency Erasing，TSE）方法，其核心思想是利用一系列有序的对抗学习器为连续帧提取互补的特征表示。具体实现上，TSE首先使用第一个学习器为第一帧提取最显著的特征。然后对于第二帧的特征图，利用时序信息擦除掉第一个学习器关注的区域。之后将移除擦除区域的特征图输入到第二个学习器中发现新的判别区域。通过递归擦除所有之前帧已经发现的区域，这些有序的学习器能够为连续帧挖掘互补的区域，最终获得目标行人的完整特征表示。如图3.1(c)所示，在TSE的驱动下，连续帧的特征可以关注到不同的身体部件，覆盖了目标行人的整个身体区域。

然而，TSE递归擦除第二帧及其后续帧的最显著区域，会损害最显著区域的视频特征表示。针对这个问题，本章提出了一个时序显著性增强（Temporal Saliency Boosting，TSB）方法，利用时序信息在视频帧之间传播最显著的信息。

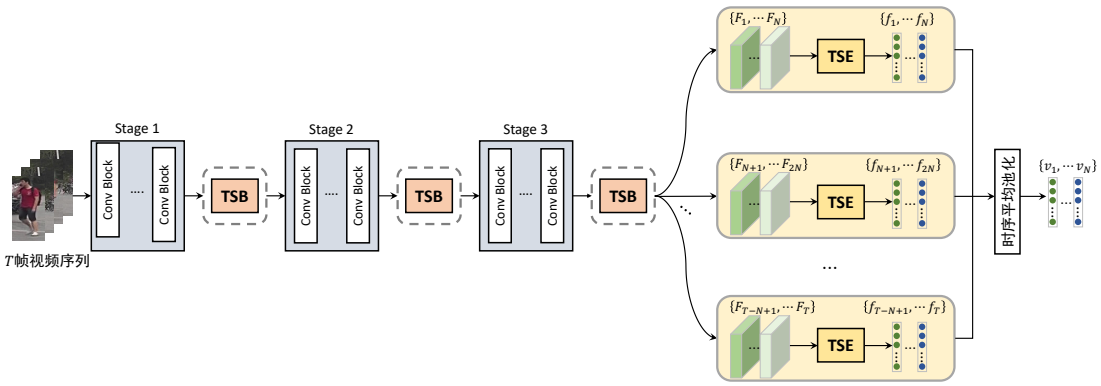


图 3.2 时序互补学习网络（Temporal Complementary Learning Network，TCLNet）结构图

通过这种方式，最显著的特征可以捕捉到所有帧的视觉线索，呈现出更强的表征能力。

本章剩余章节安排如下：章节3.2详细介绍了基于时序互补建模的特征学习方法；章节3.3给出了本章方法在公开数据集上的有效性分析；章节3.4将对本章内容进行小结。

3.2　基于时序互补建模的特征学习

本章提出一个用于视频行人重识别的时序互补学习网络TCLNet。TCLNet的结构如图3.2所示。TCLNet包含两个组件：时序显著性擦除模块TSE和时序显著性增强模块TSB。其中，TSE旨在为连续帧挖掘互补的特征表示，TSB旨在增强视频中最显著特征表示。

形式化地，给定一个包含连续T帧的视频，首先使用嵌入TSB的主干网络为每帧提取卷积特征图，定义为$\mathcal{F} = \{F_1, F_2, \ldots, F_T\}$。由于行人的判别视觉线索有限，本方法只为连续$N$（$N < T$）帧提取互补的特征表示。具体地，将$\mathcal{F} = \{F_1, F_2, \ldots, F_T\}$等分为$L$个视频片段$\{C_k\}_{k=1}^{L}$，其中每个片段包含连续$N$帧的卷积特征图，形式化为$C_k = \{F_{(k-1)N+1}, \ldots, F_{kN}\}$。然后将每个视频片段输入到TSE，为该片段的连续帧提取片段级别特征：

$$c_k = \{f_{(k-1)N+1}, \ldots, f_{kN}\} = \mathrm{TSE}(F_{(k-1)N+1}, \ldots, F_{kN}) = \mathrm{TSE}(C_k). \tag{3.1}$$

最后对$\{c_k\}_{k=1}^{L}$执行时序平均池化操作，整合所有片段的特征生成视频级别特征$\{v_1, \ldots v_N\}$。其中，v_i表示第i个学习器提取的视频特征：

$$v_i = \frac{1}{L}\sum_{k=1}^{L}\{c_k\}_i = \frac{1}{L}\sum_{k=1}^{L} f_{(k-1)N+i} \quad (1 \le i \le N). \tag{3.2}$$

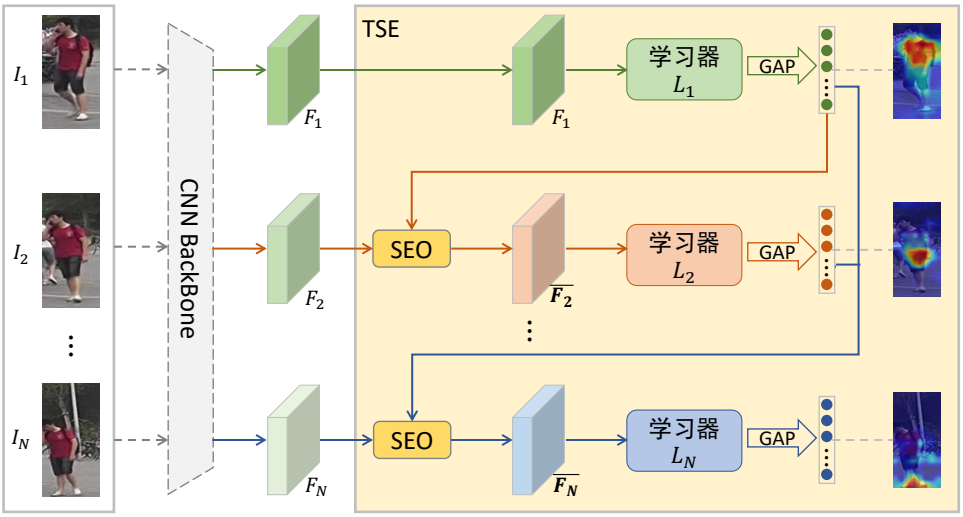


图 3.3 时序显著性擦除模块（Temporal Saliency Erasing，TSE）结构图

在测试阶段，将所有学习器提取的特征级联获得最终的视频特征v，形式化为$v = [\hat{v}_1, \ldots, \hat{v}_N]$，其中$\hat{v}_i$表示$v_i$的$L_2$归一化向量。

在接下来的章节中，分别详细介绍了时序显著性擦除模块TSE，时序显著性增强模块TSB，以及TCLNet使用的具体结构和目标函数。

3.2.1 时序显著性擦除模块

在视频行人重识别任务中，现有方法通常对每一帧执行相同的操作，导致连续帧的特征高度冗余。如图3.1所示，这些冗余的特征趋向于关注同一个局部区域，难以区分局部相似的不同行人。针对这个问题，本章设计了一个时序显著性擦除模块TSE，旨在为连续帧挖掘互补的空间区域，从而形成目标行人的一个完整特征表示。

3.2.1.1 整体结构

TSE的结构如图3.3所示。TSE对每帧迭代执行以下两个操作：一个显著性擦除操作（Saliency Erasing Operation, SEO）对抗擦除之前帧已经关注的区域；一个特定的学习器（Learner）挖掘新的互补区域。具体来说，给定一个视频片段$\{I_n\}_{n=1}^N$，利用卷积网络提取每一帧的特征，得到所有帧的卷积特征图$\{F_n \in \mathbb{R}^{H\times W\times D}\}_{n=1}^N$，其中$H$、$W$和$D$分别表示特征图的高度，宽度和通道数。然后TSE使用第一个学习器L_1为第一帧提取最显著的特征 $f_1 \in \mathbb{R}^{D_1}$：

$$f_1 = \mathrm{GAP}(L_1(F_1)), \tag{3.3}$$

其中GAP（Global Average Pooling）表示全局平均池化操作。在f_1关注的显著区域指导下，SEO显式擦除第二帧的特征图F_2，然后将擦除后的特征图送入到第二个学习器L_2中。由于L_1关注的区域已经被擦除，L_2自然地被驱使发现新的区域以识别目标行人。递归地，对于特征图$F_n\,(n>1)$，TSE首先使用SEO擦除之前帧已经关注的区域，得到一个擦除特征图$\overline{F_n}$；然后使用指定学习器L_n挖掘新的区域，获得特征向量f_n。整个过程形式化为：

$$\overline{F_n}=\mathrm{SEO}\left(F_n;f_1,\ldots,f_{n-1}\right),\quad f_n=\mathrm{GAP}\left(L_n\left(\overline{F_n}\right)\right)\ (1<n\leq N). \tag{3.4}$$

总的来说，给定一个包含连续N帧的输入片段，TSE重复执行显著性擦除操作以及使用指定学习器提取其特征向量，最后结合这N帧的特征向量获得目标行人的一个完整表征。

3.2.1.2　显著性擦除操作

显著性擦除SEO的流程如图3.3所示。SEO由关联层、区域二值化层、擦除操作三部分组成。

SEO-关联层　SEO首先设计了一个关联层，目的是获得之前帧的特征向量f_k与待擦除特征图F_n的关联图。具体实现上，首先将F_n每个空间位置的特征向量视作一个D维局部描述子$F_n^{(i,j)}$，然后使用点乘相似度 [92] 计算F_n所有局部描述子与f_k之间的语义相关性，获得一个关联图$R_{nk}\in\mathbb{R}^{H\times W}$：

$$R_{nk}^{(i,j)}=\left(F_n^{(i,j)}\right)^T\left(w^T f_k\right)\quad(1\leq i\leq H,1\leq j\leq W,1\leq k\leq n-1), \tag{3.5}$$

其中$w\in\mathbb{R}^{D_1\times D}$是一个可学习参数。从等式3.5可以推断出，描述$f_k$激活区域的局部描述子会在$R_{nk}$中呈现更高的值。因此，关联图$R_{nk}$可以在$F_n$中定位到第$k$帧激活的区域。

SEO-区域二值化层　在关联层之上，SEO设计了一个区域二值化层，目的是基于关联图生成一个二值掩码以定位待擦除区域。最简单的实现方式是在关联图上设置一个阈值，将高于阈值的特征单元进行擦除。但是这种方式会产生一个离散的擦除区域。由于卷积特征单元具有空间相关性，当不连续地擦除特征单元时，擦除单元的信息依旧可以传输到下一个卷积层 [105]。基于这个现象，本方法提出擦除特征图的一个连续区域，从而完全抹去擦除区域的信息。

如图3.4所示，区域二值化层使用一个滑动窗口在关联图中搜索一个最显著的连续区域。具体实现上，给定一个$H\times W$的关联图和一个$h_e\times w_e$的滑动窗口，

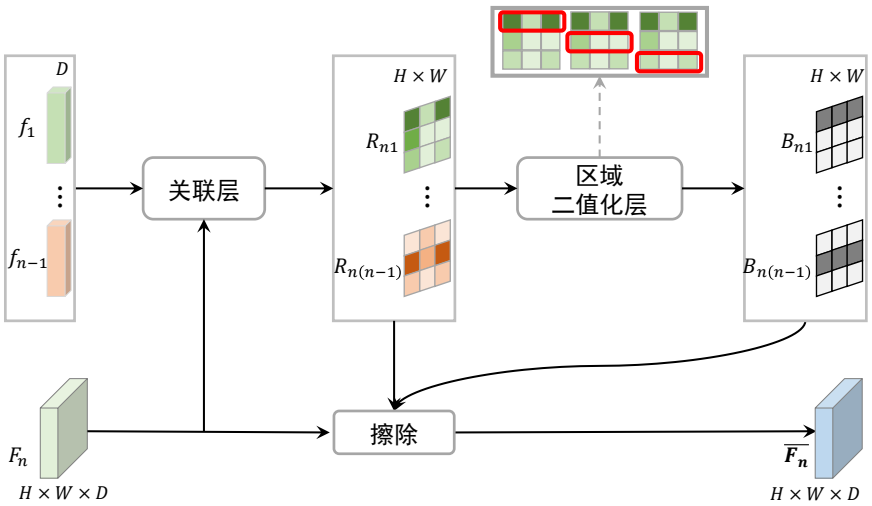


图 3.4 TSE的显著性擦除操作（Saliency Erasing Operation, SEO）流程图

分别以垂直步长s_h和水平步长s_w移动这个滑动窗口，此时滑动窗口所在的位置总数为$N_{pos} = \left(\lfloor \frac{H-h_e}{s_h} \rfloor + 1\right) \times \left(\lfloor \frac{W-w_e}{s_w} \rfloor + 1\right)$，因而每个关联图有$N_{pos}$个候选区域；然后，每个候选区域的关联值设为区域内所有单元关联值的总和，选择关联值最高的候选区域作为待擦除区域；最后，给定第n帧的关联图R_{nk}，将擦除区域内单元的值设为1，生成对应的二值掩码$B_{nk} \in \mathbb{R}^{H \times W}$。之后合并第$n$帧的所有二值掩码为$F_n$生成最终掩码$B_n$：

$$B_n = B_{n1} \odot B_{n2} \odot \cdots \odot B_{n(n-1)}, \tag{3.6}$$

其中$\odot$是逐元素乘积运算。

SEO-擦除操作　为了使TSE可以通过梯度反传算法进行优化，TSE采用一个门控机制对特征图F_n执行擦除操作。实现上，TSE使用Softmax函数对关联图进行归一化，然后使用B_n擦除选中的区域，获得一个门控图$G_n \in \mathbb{R}^{H \times W}$：

$$G_n = \text{Softmax}\left(R_{n1} \odot R_{n2} \odot \cdots \odot R_{n(n-1)}\right) \odot B_n. \tag{3.7}$$

最后基于G_n擦除F_n，生成擦除特征图$\overline{F_n}$。由于之前帧关注的区域已被移除，学习器L_n能够对$\overline{F_n}$挖掘新的互补区域，从而提取到一个更完整的行人视频特征。

3.2.2　时序显著性增强模块

尽管TSE能够为连续帧提取互补的特征，但是显著性擦除操作不可避免地会造成最显著区域信息的损失。针对这个问题，本章设计了一个显著性增强模块TSB。在擦除操作之前，视频帧的中间层卷积特征会集中关注最显著的区域。

基于这个现象，TSB提出在所有帧的中间层卷积特征之间传播最显著的信息。通过这种方式，最显著区域的特征能够捕捉到所有帧的视觉线索，呈现出更强的表现力。

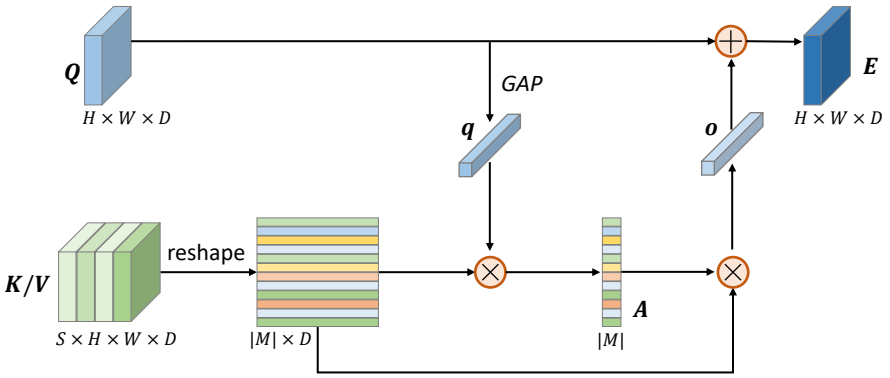


图 3.5 时序显著性增强模块（Temporal Saliency Boosting，TSB）结构图

TSB的结构如图3.5所示。TSB基于自注意力（Self-attention）[26] 机制实现，将每帧的特征图作为查询$Q \in \mathbb{R}^{H\times W\times D}$；输入视频中剩余$S$帧的特征作为键$K \in \mathbb{R}^{S\times H\times W\times D}$和值$V \in \mathbb{R}^{S\times H\times W\times D}$。具体实现上，首先将$Q$压缩为一个可以描述查询统计信息的向量。本方法采用全局平均池化操作生成一个通道统计向量$q \in \mathbb{R}^{D}$；然后将K和V调整为$|M| \times D$（$|M| = S \times H \times W$）大小。此时$K$和$V$可以视为$|M|$个键向量和值向量的集合；之后计算查询向量和每个键向量的余弦相似性，得到所有值向量的权重$A \in \mathbb{R}^{|M|}$：

$$A_i = \frac{\exp(\tau \bar{q}^T \bar{K}_i)}{\sum_{j=1}^{|M|} \exp(\tau \bar{q}^T \bar{K}_j)}, \tag{3.8}$$

其中$K_i \in \mathbb{R}^{D}$表示K的第i个键向量，$\bar{q}$和$\bar{K}_i$分别是q和K_i的L_2归一化向量，τ是控制权重分布的超参数。之后根据权重系数对所有值向量进行加权求和，获得输出特征$o \in \mathbb{R}^{D}$，形式化为$o = V^T A$。

最后，使用残差学习机制将o添加到查询特征Q，形式化为$E = \mathrm{BN}(o) + Q$。其中BN表示批量归一化层（Batch Normalization）[106]，用来调整o相对于Q的大小。值得注意的是，$o \in \mathbb{R}^{D}$在进入BN层之前，首先沿着空间维度扩展到$H \times W \times D$大小，以与Q的大小兼容。

3.2.3　网络结构和目标函数

网络结构　TCLNet使用ResNet50 [85] 作为主干网络。ResNet50由四个卷积阶段Stage1~4组成，四个卷积阶段分别包含3，4，6和3个残差块。本章使用前三

个卷积阶段作为主干网络，最后一个卷积阶段作为TSE的学习器。为了降低模型复杂度，TSE的多个学习器共享前两个残余块的参数，最后一个残余块具有独立参数。TSB可以嵌入到主干网络的任意阶段，TSE添加在主干网络的末端。

目标函数 遵循标准身份分类范式 [14]，本章对每个视频向量v_i（等式3.2）添加一个分类层，并且使用交叉熵损失指导对应学习器的训练。最近的工作 [49, 68, 97] 结合交叉熵损失和困难样本三元组 [20] 损失训练网络。为了与这些方法公平比较，本章还探索了在训练期间为每个视频向量使用困难样本三元组损失。交叉熵损失和困难样本三元组损失的详细介绍见章节2.2.3.1。

3.3 实验分析

本节分别在MARS [43] 、DukeMTMC-VideoReID [76] 和LS-VID [94] 三个视频数据集上对本章提出的时序互补建模方法进行评测。本节接下来分别介绍了实验细节、消融实验、与当前先进方法的对比实验、与第二章方法的对比和结合、在图像行人重识别的拓展以及可视化分析。

3.3.1 实现细节

本章使用Adam优化器进行网络优化，共训练150个回合。初始化学习率设为3×10^{-4}，批量大小为32,，每训练40个回合后下降10倍。输入图像的大小为256×128。对于每个原始视频，以8帧为间隔随机采样4帧形成一个片段。本章使用水平翻转和随机裁剪作为数据增广。TSB嵌入到主干网络的第二个卷积阶段。在TSE中，学习器的数量N设为2，擦除区域的长度h_e设为3，宽度w_e设为8。三元组损失的距离阈值设置为0.4。

本文使用视频的全部帧进行测试。首先将每个原始测试视频等分为多个包含4帧的视频片段，然后利用训练完成的TCLNet提取每个视频片段的特征。最后将所有视频片段的特征取平均得到原始测试视频的特征。对于每个查询视频，计算它与所有候选视频特征的余弦相似值。根据计算出的特征相似值对所有的候选视频进行排序，选出最相似的候选视频，达到目标检索的目的。

3.3.2 消融实验

为了验证TCLNet各个模块的有效性，本节在MARS数据集上采取了一系列消融实验。本节首先引入了一个常用的基准模型（Baseline），使用ResNet50独

表 3.1 TCLNet在MARS数据集上的消融实验。GFLOPs表示模型处理一帧片段所需的浮点运算次数，Param.表示模型的参数量

方法	GFLOPs	Params	mAP (%)	top-1 (%)
Baseline	4.061	23.5M	79.6	86.8
Baseline+TSE-wo-SEO	4.061	27.9M	80.9	87.2
Baseline+TSE	4.062	29.9M	**82.5**	**88.2**
Baseline+TSB	4.063	23.5M	82.3	87.6
TCLNet (TSE+TSB)	4.065	29.9M	**83.0**	**88.8**
TCLNet (TSE+TSB-stage1)	4.067	29.9M	82.2	87.4
TCLNet (TSE+TSB-stage2)	4.065	29.9M	**83.0**	**88.8**
TCLNet (TSE+TSB-stage3)	4.064	29.9M	82.6	88.2
TCLNet (TSE+TSB-stage23)	4.066	29.9M	82.7	88.2

立提取每一帧的特征，然后通过时序平均池化融合多帧特征得到一个视频特征。为了公平比较，本节所有的模型都使用交叉熵损失进行训练，并采用与TCLNet相同的训练细节。

显著性擦除模块TSE的有效性　本节测试了TSE的有效性，引入了一个比较模型Baseline+TSE。Baseline+TSE将Baseline的stage4卷积阶段替换为TSE。如表3.1所示，TSE显著提升了重识别精度，在mAP和top-1上提升了2.9%和1.4%，且只增加了不到1%的计算开销。这是由于TSE的一系列学习器协同工作对连续帧挖掘互补的行人区域，能够生成一个目标行人的整体表征，有助于区分非常相似的不同行人。

TSE的显著性擦除操作SEO的有效性　为了验证TSE性能的提升不仅仅是来自学习器参数的增加，本节引入了一个变体TSE-wo-SEO。TSE-wo-SEO对连续帧采用一系列有序学习器，但是没有使用显著性擦除操作SEO。如表3.1所示，相比于Baseline，TSE-wo-SEO只带来了很少的性能提升。这是由于在没有显著性擦除操作下，不同学习器捕捉到的视觉线索几乎相同。从表3.1中可以观察到，TSE的性能显著优于TSE-wo-SEO，在mAP和top-1上分别提升了1.6%和1.0%，验证了SEO的有效性。性能的提升归功于SEO引导不同的学习器关注不同的行人区域，有利于发现完整的视觉线索。综上分析，TSE的性能提升主要来自显著性擦除操作而不是增加的参数。

表 3.2 学习器个数对显著性擦除模块TSE性能的影响

学习器个数	GFLOPs	Params	mAP (%)	top-1 (%)
1 (Baseline)	4.061	23.5M	79.6	86.8
2	4.062	29.9M	**82.5**	88.2
3	4.063	34.5M	82.4	**88.4**
4	4.063	38.9M	81.0	87.0

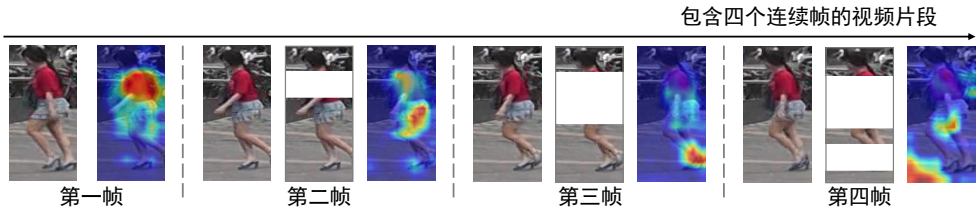


图 3.6 Baseline+TSE模型特征图的可视化

学习器个数的影响 TSE包含N个有序的学习器，旨在为N个连续帧挖掘互补区域。表3.2研究了学习器个数对模型Baseline+TSE性能的影响。可以观察到，随着使用更多的学习器去挖掘互补区域，模型性能也会随之提升。但是，当N达到4时，重识别性能会大幅下降。这是由于当N过大时，输入片段最后一帧的大部分显著区域都被擦除，此时最后一个学习器会激活背景等非判别区域，破坏了最终的视频特征表示。一个示例如图3.6所示。另外，从表3.2可以观察到与$N = 2$相比，$N = 3$并没有带来显著的性能增益。考虑到模型的复杂性，本章将N设置为2。

擦除区域大小的影响 本章使用256×128作为输入分辨率，ResNet50的前三个stage作为主干网络，可以计算出TSE的输入卷积特征图大小为16×8。由于行人的空间结构性，本章将擦除区域宽度固定为8，以对特征图的整行进行擦除。表3.3研究了擦除区域高度对TSE性能的影响。可以观察到，当擦除区域高度为3时，TSE取得最高的性能水平。擦除区域高度更大或更小都会降低重识别性能。这是由于：（1）过小的擦除区域无法有效驱动当前学习器发现互补区域；（2）过大的擦除区域会迫使当前学习器激活背景等非判别区域。因此，本方法将擦除区域高度设置为3。

显著性增强模块TSB的有效性 本节评估了TSB模块的有效性。本节引入了一个比较模型Baseline+TSB，该模型将TSB添加到Baseline的stage2。如表3.1所

表 3.3 擦除区域高度对显著性擦除模块TSE性能的影响

擦除区域高度	GFLOPs	Params	mAP (%)	top-1 (%)
2	4.062	29.9M	81.8	87.2
3	4.062	29.9M	**82.5**	**88.2**
4	4.062	29.9M	82.0	87.4
5	4.062	29.9M	81.7	87.2

示，TSB单独带来了2.7%的mAP和0.8%的top-1增益，并且只增加了不到1%的计算开销。实验表明了TSB能够增强特征表示能力。从表3.1可以观察到：TCLNet进一步提高了大约1%的mAP和top-1。性能的进一步提升验证了TSE和TSB是功能互补的。具体来说，TSE擦除最显著区域以挖掘更多的判别线索；TSB在视频帧之间传播最显著信息以增强最显著特征表示。通过这种方式，TSB缓解了TSE擦除操作造成的显著性信息损失，提取到与TSE互补的特征表示。

TSB模块放置位置的影响　表3.1比较了放置在ResNet50不同阶段的TSB性能。可以观察到，放置在stage2和stage3的TSB模块更为有效。一个可能的解释是stage1的卷积特征图表现力较弱，不足以提供精确的语义线索。本节还展示了放置更多TSB模块的结果，即分别在主干网络的stage2和stage3添加一个TSB模块。如表3.1所示，添加更多的TSB模块并不会带来性能增益。因此本方法只在主干网络的stage2嵌入一个TSB模块用于视频行人重识别。

计算复杂度分析　为了说明TSB和TSE的计算开销，表3.1报告了模型平均处理一帧所需的浮点运算次数（GFLOPs）以及模型的参数量（Params）。可以观察到，TSE和TSB模块增加了非常少的计算开销。TCLNet平均处理一帧需要4.065GFLOPs，相比于Baseline只增加了0.08%。增加的计算量主要来自TSE的关联图和TSB的权重概率。由于这两个操作可以通过矩阵乘法实现，TSE和TSB模块在GPU上只占用很少的时间。TCLNet引入了6.4%的参数量，增加的参数主要来自TSE的多个学习器。值得注意的是，仅仅使用多个学习器（Baseline+TES-wo-SEO）只能带来不到1%的mAP和top-1提升，表明TCLNet的性能提升不只是由于参数量的增加。

表 3.4 TSE与其他擦除方法在MARS数据集上的对比

方法	GFLOPs	Params	mAP (%)	top-1 (%)
Baseline	4.061	23.5M	79.6	86.8
Baseline+DropBlock [105]	4.061	23.5M	79.8	86.9
Baseline+RE [107]	4.061	23.5M	81.5	86.6
Baseline+TSE	4.062	29.9M	**82.5**	**88.2**
Baseline+TSE+RE	4.062	29.9M	**82.9**	88.1

表 3.5 TSB与其他特征传播方法在MARS数据集上的对比

方法	GFLOPs	Params	mAP (%)	top-1 (%)
Baseline+3D卷积 [64]	5.689	33.7M	80.0	86.1
Baseline+NonLocal [92]	5.403	25.6M	**82.4**	**87.6**
Baseline+TSB	4.064	23.5M	82.3	**87.6**
Baseline+TSE+NonLocal	5.405	32.1M	82.6	87.6
Baseline+TSE+TSB (TCLNet)	4.065	29.9M	**83.0**	**88.8**

3.3.3 与相似方法的比较

比较TSE和其他擦除策略 表3.4在MARS数据集上比较了TSE和其他擦除方法。对比的方法包括DropBlock [105] 和RE [107]。其中，DropBlock在训练期间随机丢弃卷积特征的一个连续区域，目的是减少模型过拟合的风险。而TSE同时应用在训练和测试阶段，在前一帧激活区域的指导下丢弃当前帧的一个连续区域，目的是提取连续帧的互补特征。如表3.4所示，DropBlock只带来了不到0.3%的性能提升，而TSE的性能显著优于DropBlock，在mAP和top-1上分别提升了2.7%和1.3%。实验结果表明本章提出的擦除策略对视频行人重识别任务是更有效的。 RE [107] 在训练期间随机擦除输入图像的一个矩形区域，是一种广泛使用的数据增强技术。如表3.4所示，TSE高于RE方法1%的mAP和1.6%的top-1。作为一种数据增强技术，RE可以与本章方法结合，进一步提升0.4%的mAP。

比较TSB和其他特征传播策略 表3.5在MARS数据集上比较了TSB和其他特征传播方法。对比方法包括3D卷积 [64] 和NonLocal [92]。对于3D卷积，本节采用基于ResNet50的3D网络，在ResNet50架构中使用3D卷积核提取时空特征。如表3.5所示，TSB的性能显著高于3D卷积。这是由于3D卷积参数巨大而难以优

表 3.6 本章方法与现有视频行人重识别方法在MARS、DukeMTMC-Video和LS-VID数据集上的性能对比

方法		MARS		Duke-Video		LS-VID	
		mAP	top-1	mAP	top-1	mAP	top-1
集合建模	STAN [46]	65.8	82.3	-	-	-	-
	EUG [76]	67.4	80.8	78.3	83.6	-	-
	RQEN [45]	71.7	77.8	-	-	-	-
	TAFD [47]	78.2	87.0	-	-	-	-
时空建模	M3D [52]	74.1	84.4	-	-	40.1	57.7
	Snippet+OF [61]	76.1	86.3	-	-	-	-
	V3D [49]	77.0	84.3	-	-	-	-
	GLTP [94]	78.5	87.0	93.7	96.3	44.3	63.1
	CoSAM [97]	79.9	84.9	93.7	96.2	-	-
	IANet (第二章) [68]	85.0	**90.2**	96.1	**96.9**	69.2	80.1
互补建模	TCLNet（本章方法）	**85.1**	89.8	**96.2**	**96.9**	**70.3**	**81.5**
结合第二章方法	TCLNet+IA	**85.5**	**90.3**	**96.6**	**97.2**	**72.8**	**83.7**

化，很容易过拟合到训练数据，并且受限于短期时序的特征传播。而TSB模块能够以更少的参数量捕捉长期的时序依赖。相比于NonLocal [92]，TSB能够以更少的计算开销和参数量实现与之相当的性能。更重要的是，TSE+TSB的性能显著优于TSE+NonLocal，说明了TSB能够与TSE更好地结合。这是由于TSB只在视频帧之间传播最显著信息，与TSE的功能更为互补。

3.3.4　与当前先进方法的对比

本节将TCLNet与当前先进视频行人重识别方法进行比较。为了对比公平，本节使用交叉熵损失和困难三元组损失联合训练TCLNet。如表3.6所示，在三个视频数据集上，本文提出的TCLNet与众多模型相比均达到了较高水平。相比于集合建模的方法 [45–47, 76]，TCLNet能够取得更优的性能，验证了时序信息对于视频行人重识别任务的有效性；最近工作M3D和V3D [49, 52] 使用三维卷积以及NonLocal建模视频的时序信息，但是这两类方法都有较高的计算开销。TCLNet具有更少的计算开销，同时取得了更优的性能。性能的提升归功于TCLNet从连续视频帧中学习到互补特征，增强了视频特征的表示完整性。

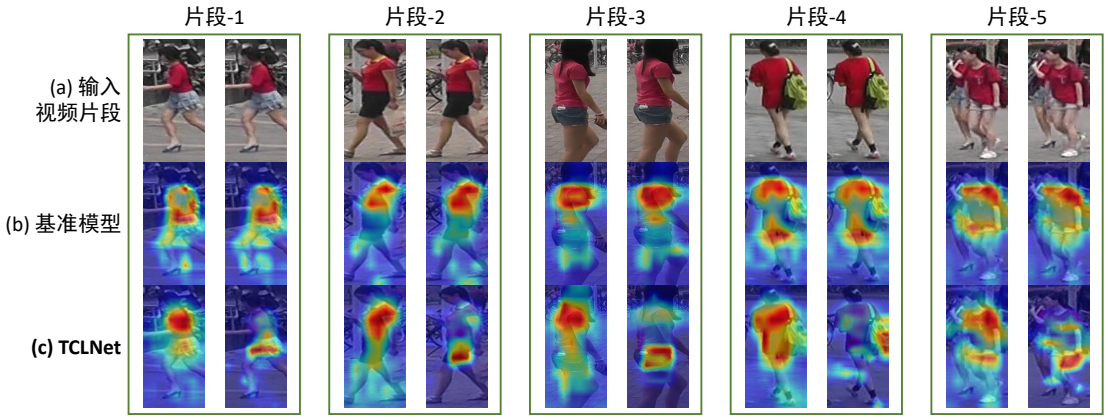


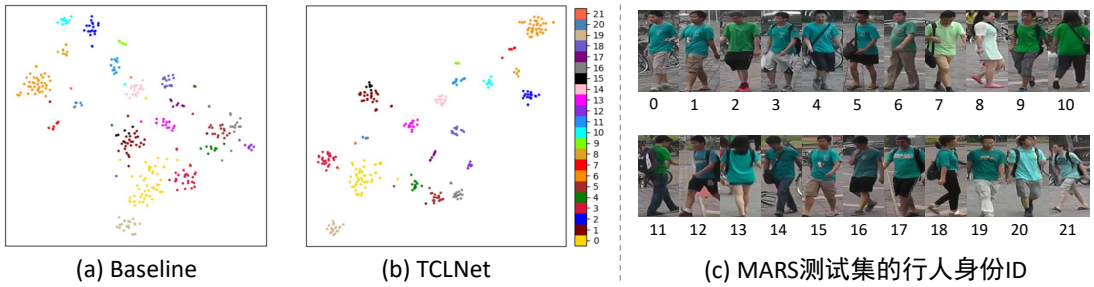
图 3.7 基准模型（Baseline）和本文方法（TCLNet）特征图的可视化

3.3.5 与第二章方法IA的对比和结合

本节详细比较了TCLNet和第二章的IANet。从表3.6中可以观察到：（1）在数据集MARS和DukeMTMC-VideoReID上，TCLNet和IANet达到了接近的性能。然而IANet需要借助行人分割模型指导部件特征的提取，而TCLNet无需借助外部姿态信息，可以自适应在不同帧中挖掘判别的身体区域。（2）TCLNet在LS-VID数据集上取得了更高的精度。这是由于LS-VID存在大量衣物非常相似的不同行人，TCLNet通过互补建模得到一个更完整的行人表征，有利于区分这些极其相似的不同行人，从而取得更高的精度。（3）TCLNet与IA模块功能互补，结合IA能够进一步提高重识别精度，在LS-VID数据集上带来了2.5%mAP和2.2%top-1的提升。实验结果验证了TCLNet和IA是功能互补的，结合两种方式能够进一步提高特征的完备性。

3.3.6 基于图像的行人重识别

本章的TSE方法很容易拓展到图像行人重识别任务，其中不同的学习器为输入图像提取互补的视觉线索。本节在图像数据集Market1501 [29] 上比较了一个以ResNet50为特征提取器的基准模型Baseline，以及一个将ResNet50的第四个卷积阶段替换为TSE的Baseline+TSE模型。这两个模型都只使用交叉熵损失进行训练。Baseline和Baseline+TSE模型分别取得了80.3%和85.3%的mAP。相比于Baseline，Baseline+TSE提高了3.2%的mAP，验证了TSE的通用性和泛化性。



(a) Baseline　(b) TCLNet　(c) MARS测试集的行人身份ID

图 3.8 基准模型（Baseline）和本文方法（TCLNet）特征分布的tSNE可视化

3.3.7　可视化分析

特征图的可视化分析　本节比较了Baseline和TCLNet特征图的可视化结果。如图3.7(b)所示，Baseline的特征只集中到一个局部区域（红色上衣），很难区分图中这些都穿红色上衣的不同行人。相反，TCLNet对不同帧挖掘互补的视觉线索。如图3.7(c)所示，对于包含连续两帧的视频片段，学习器L_1提取的第一帧特征与红色上衣最为相关，而学习器L_2激活了第二帧的下半身区域。利用学习器L_2挖掘的互补区域，TCLNet能够区分具有相似外观的不同行人，显著提升重识别性能。

特征分布的可视化分析　本节从MARS测试集中选取多个具有相似外观的行人，并使用t-SNE [108] 可视化这些行人的特征分布。如图3.8所示，这些行人身穿蓝色衬衫，具有很小的类间差异。由于Baseline只关注一个局部身体部件，这些不同身份的特征互相交错且难以区分。而通过TSE提取互补特征，TCLNet提取的不同身份特征更具可分性。通过比较图3.8(a)和(b)，TCLNet能够区分Baseline难以区分的行人，例如第1个、第3个和第4个身份行人。

3.4　本章小结

本章独创性的观察到已有视频重识别方法会对连续帧产生冗余的特征，并且这些冗余的特征趋向于关注同一个显著但是局部的区域，导致最终的视频特征难以区分类间差异小的行人。根据这个发现，本章提出了一种时序建模的新思路，时序互补建模——利用时序信息对连续帧提取互补的特征，使视频特征能够覆盖一个更完整的身体区域。基于时序互补建模的思想，本章提出了一个时序互补学习网络TCLNet，在多个数据集的实验表明TCLNet有效提高了视频特征的完整性，显著提升了重识别性能，验证了时序互补建模思想对行人重识别

任务的有效性。

本章的“时序互补建模”能够与上一章的“时空交互建模”有效结合。一方面，“时空交互建模”通过在不同身体部件间传播特征，使每个部件特征融合了视频全局信息，提高了每个局部区域特征的判别力；另一方面，“时序互补建模”通过约束不同帧激活不同局部区域，使多帧能够联合覆盖更多的身体区域，提高了视频特征的完整性。在多个数据集的实验表明TCLNet结合IA（TCLNet+IA）能够进一步提高重识别精度，特别是在类间相似问题严重的LS-VID数据集上。

结合时空交互和时序互补建模在特征学习上仍然具有一定的局限性，主要体现在两个方面：1）TCLNet+IA集中在行人的表观建模上，无法有效编码行人的运动信息，导致在长时间行人重识别问题上可能失效。具体来说，在长时间重识别场景，行人很有可能更换衣服，此时衣服颜色等表观信息会失效，造成时空交互和时序互补特征学习方法的性能显著下降。2）由于TCLNet+IA以一个高分辨率处理视频序列的所有帧，计算效率比较低，可能无法达到实时场景的需求。特别地，TCLNet需要使用多个有序的学习器，无法并行处理连续帧，进一步增加了时间开销。在未来的工作中，如何在时空交互和时序互补特征学习中编码动作信息，以及提高其计算效率，是一个值得研究的问题。

第4章　面向遮挡场景的图像时空补全方法

前两章从视频行人重识别系统的**准确性**出发，针对类间相似问题提出了两个时空建模方法，有效提高了视频特征的完备性。本章聚焦在特定遮挡场景下视频行人重识别系统的**鲁棒性**。针对遮挡问题，提出了一个图像时空补全学习方法。通过设计合理的网络结构和损失函数，使模型能够修复遮挡区域的内容，提高了视频特征对遮挡的鲁棒性。

4.1　引言

经过近几年的发展，无/少遮挡场景下的行人重识别技术已逐渐趋于成熟，在遮挡较少场景下采集的数据集上已经取得了非常高的识别性能。随着应用需求的增长，由于人群密集的公共场景中采集的数据往往存在各种各样的遮挡，遮挡场景下的行人重识别引起了越来越多研究者的关注。遮挡物具有丰富多变的形状，大小、并且可能覆盖行人的任意位置，使得行人的表观出现剧烈的变化，从而增加了识别的难度，但另一方面也给行人重识别研究提供了新的机遇和方向。

最初专门解决遮挡问题的行人重识别研究集中在图像数据上 [109, 110]。已有方法大致分为两类：基于区域对齐的方法 [35–37, 111] 和基于姿态对齐的方法 [22, 32]。VPM（Visibility-aware Part Model）[40] 是区域对齐的一个代表性工作。VPM将人体划分为上、中、下三个区域，并通过自监督学习感知匹配图像之间的共享区域，然后只利用共享的局部区域计算相似性。PFGA（Pose-Guided Feature Alignment）[22] 是姿态对齐的一个代表性工作。PFGA使用姿态预测器预测人体的关节点坐标，辅助局部部件的特征生成，之后计算不丢失关节点的未遮挡部件之间的相似度。以上方法的主要思路是让网络更关注非遮挡部分的行人特征，在一定程度上提高了对遮挡问题的处理能力。但是这两类方法都采用丢弃遮挡部件的策略，无法处理非遮挡部件具有相似外观而遮挡部件为关键判别部分的情况。

在遮挡严重的场景下，单帧图像所包含的信息是非常有限的。如图4.1(a)所示，图像被其他人或静态障碍物严重遮挡，导致目标行人只有极少的可视区域。

（a）遮挡图像　　（b）遮挡序列

图 4.1 遮挡场景下行人图像和视频示例

行人身体信息的严重丢失使基于图像的遮挡行人重识别算法难以取得令人满意的性能。比如，在具有严重遮挡问题的图像数据集Occluded-DukeMTMC [22] 上，先进的PFGA方法仅能达到51.4%top-1和37.3%mAP。与图像相比，视频包含更多的可利用信息，包括行人的多种外观和运动等。如图4.1(b)所示，遮挡物只出现在一个序列的中间几帧，而其余帧包含了遮挡帧中丢失的行人腿部信息。因此，视频数据从本质上更容易在遮挡条件下识别行人，基于视频的遮挡行人重识别具有重要的研究意义。

针对行人视频序列的遮挡问题，已有方法借助于时间注意力机制 [44, 58, 59]。实现上，在进行特征学习的同时，对序列中每一帧的质量进行评估，并使用估计的质量分数对所有帧特征进行加权融合。基于时间注意力的方法能够避免低质量的遮挡帧参与视频特征融合，从而降低了遮挡帧的不利影响。然而，这类方法存在以下两个问题：（1）如图4.1(b)所示，遮挡帧中可视的身体区域依旧可以提供强有力的判别线索，因此直接丢弃遮挡帧会造成行人外观的损失；（2）丢弃的遮挡帧中断了视频序列的时序性，使模型无法利用步态等时序信息辅助重识别。因此，已有方法会造成遮挡行人序列外观信息和时序信息的损失，难以处理遮挡部件为判别因素的情况，使遮挡场景下的视频行人重识别任务依旧极具挑战性。

基于上述分析，本章认为相比于直接丢弃掉遮挡帧，更为科学的方案是利用时空信息修复遮挡区域的内容，从而保留行人完整的外观和动作信息。为了实现上述目标，本章设计了一个图像层面的时空补全网络（Spatial-Temporal Completion network，STCnet），旨在利用图像生成技术复原遮挡帧的像素。通过结合时空补全网络和已有的视频行人重识别模型，本章提出了一个遮挡鲁棒视频行人重识别框架VRSTC（Video reID framework occlusion Robust with Spatial-Temporal Completion）。实验表明，本章方法有效提高了对遮挡行人的重识别准确率。

本章剩余章节安排如下：章节4.2主要介绍本章所提出的时空补全网络；章节4.3将具体介绍本章所构建的遮挡鲁棒视频行人重识别框架；章节4.4将具体介绍Occluded-DukeMTMC-VideoReID数据集的构建过程并对其进行统计分析；章节4.5给出本章方法在Occluded-DukeMTMC-VideoReID数据集上以及其他公开数据集的评测结果；章节4.6将对本章内容进行小结。

4.2　时空补全网络

由于人体具有空间拓扑结构以及行人序列的帧与帧之间存在时间相关性，当某一帧发生部分遮挡时，我们很容易通过序列中的可视区域推断出遮挡区域的内容。受到这一启发，本章设计了一个时空补全网络，旨在利用时空信息复原遮挡区域的行人外观，从而降低遮挡的不利影响。

4.2.1　网络结构

如图4.2所示，时空补全网络STCnet由四个部件组成：空间生成器，时序生成器，判别器和身份保持器。空间生成器利用人体的拓扑结构，根据当前帧的可视部件初步预测出遮挡区域的像素。时序生成器利用帧与帧的时间关联，根据相邻帧的信息对初步预测的结果进行精细化。STCnet的目标是生成器能够将输入的遮挡帧复原为相应的未遮挡且身份不变的真实帧。为了达到这个目的，STCnet引入了判别器，使用对抗学习保证生成帧的真实性。与此同时，STCnet还引入了身份保持器来保证生成帧具有正确的身份信息。

空间生成器　根据人体的空间拓扑结构，遮挡区域的内容可以根据当前帧的可视区域推测出来。本章设计了一个空间生成器（Spatial Generator），通过堆叠卷积操作隐式建模遮挡区域与可视区域的空间关联，使空间生成器能够利用可见的身体部件复原出被遮挡的部件。空间生成器采用编码器-解码器 [112] 结构。编码器由多个卷积层堆叠而成，以遮挡帧作为输入，输出一个融入空间上下文的隐编码表示。解码器由多个转置卷积层（Transpose Convolution） [113] 堆叠而成，以编码器生成的隐编码作为输入，输出被遮挡区域的像素值。此外，编码器采用空洞卷积（Diated Convolution） [114] 以扩大感受野大小，有利于将远距离的可视部件内容传播到遮挡区域。

空间生成器采用与文献 [112] 相同的网络结构。编码器由5个卷积层和4个空洞卷积层堆叠而成，解码器由2个转置卷积层堆叠而成。所有卷阶层都采用3 ×

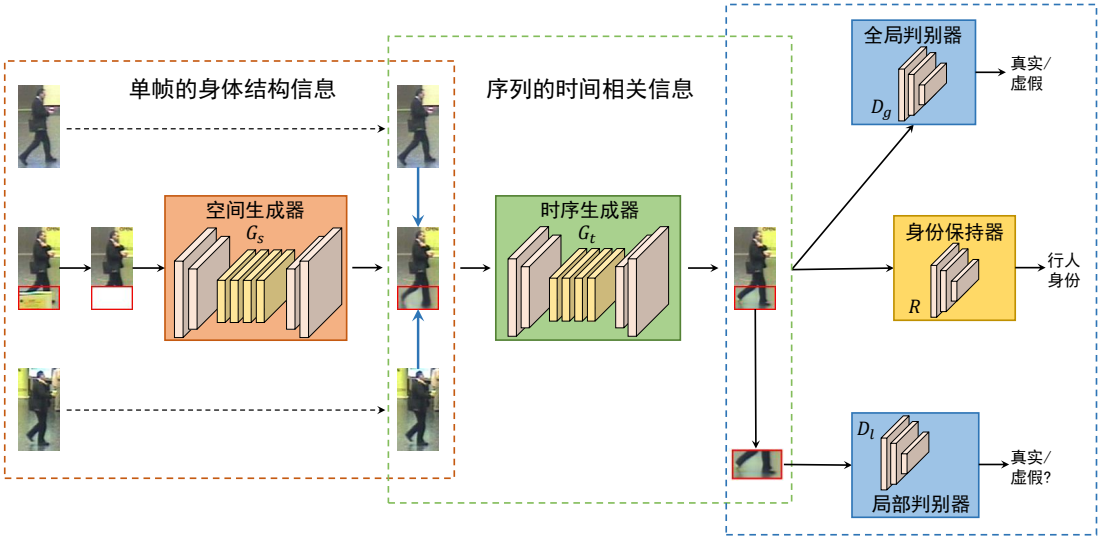


图 4.2 时空补全网络（Spatial-Temporal Completion network，STCnet）的结构图

3的卷积核和ELUs [115] 作为激活函数。此外，空间生成器在编码器和解码器之间采用类似U-Net [116] 的跨层连接，这种跨层连接在图像生成任务 [117] 能够取得较好的效果。

时序生成器 考虑到帧与帧之间的时间相关性，相邻帧的信息也能帮助修复遮挡区域的内容。本章设计了一个时序生成器（Temporal Generator），依据相邻帧的信息对空间生成器输出的内容进行精细化。

图4.3展示了时序生成器的结构设计。时序生成器包含编码器、时序注意力层、解码器三个部分。编码器的目的是为空间生成器的输出（称作“当前帧”）以及相邻前后帧提取卷积特征；时序注意力层的目的是在相邻帧的卷积特征图中定位到与遮挡部件高度相关的区域，从而将相邻帧的关联信息传递到遮挡帧中。具体实现上，由于具有相同外观的特征单元一般具有较高的相似性[92]，时序注意力层根据帧间特征单元的相似性实现遮挡部件在相邻帧的定位。如图4.3所示，给定当前帧的编码特征F和相邻帧的编码特征R^1/R^2，其中R^1表示F前一帧的编码特征，R^2表示F后一帧的编码特征。本节以R^1为例说明时序注意力层的具体操作。首先将F和R^1划分为多个3×3的局部区域，计算F和R^1任意两个局部区域的余弦相似性，并使用softmax函数对结果进行归一化：

$$s^1_{a,b,a',b'} = softmax\langle \frac{f_{a,b}}{||f_{a,b}||_2}, \frac{r^1_{a',b'}}{||r^1_{a',b'}||_2} \rangle, \tag{4.1}$$

其中，$f_{a,b}$表示F的中心点位于(a,b)的局部区域的特征，$r^1_{a',b'}$表示R^1的中心点位于(a',b')的局部区域的特征，$s^1_{a,b,a',b'}$表示$f_{a,b}$和$r^1_{a',b'}$的相似值。然后对每一个位

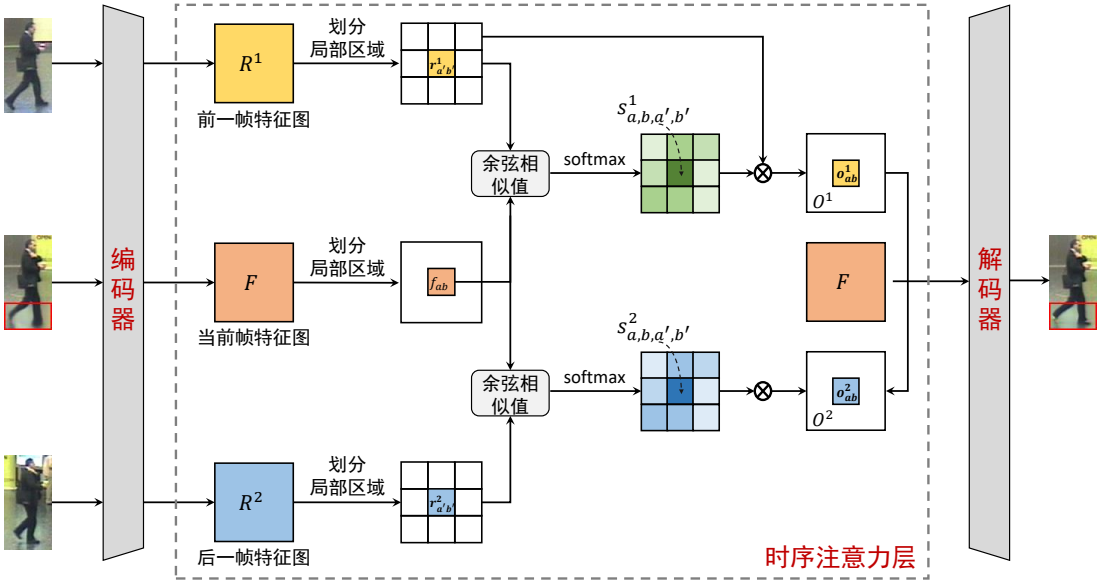


图 4.3 时空补全网络中时序生成器的结构图

置的局部区域，根据相似值对R^1所有区域特征进行加权求和，获得更新后的区域特征：

$$o^1_{a,b} = \sum_{a'b'} s^1_{a,b,a',b'} r^1(a',b'). \tag{4.2}$$

最后根据区域位置(a,b)排列组合$\{o^1_{a,b}\}_{a,b}$，得到更新后的卷积特征O^1。按照上述操作，将F和R^2输入到时序注意力层，得到更新后的卷积特征O^2。通过上述的时空注意力机制，O^1和O^2分别融合了前一帧和后一帧中与遮挡部件高度关联的信息，从而可以帮助遮挡区域的复原；最后，将编码器特征F和时间注意力层生成的特征$\{O^1, O^2\}$输入到解码器中，联合预测遮挡区域的像素。时序生成器的编码器与解码器采用与空间生成器相同的网络结构。

判别器　为了保证生成帧的真实性，STCnet引入了生成器和判别器之间的对抗学习 [118]。STCnet使用两个判别器：局部判别器（Local Discriminator）和全局判别器（Global Discriminator）。局部判别器以遮挡区域对应的生成作为输入，判断生成器输出的内容是否真实。局部判别器有利于生成身体部件的细节信息。全局判别性以生成的整帧作为输入，约束生成帧的全局结构性。这两个判别器协同工作，不仅保证了生成内容的真实性，而且确保它与上下文身体信息保持一致。

局部判别器和全局判别器采用相同的网络结构，都由6个卷积层堆叠并最终经过一个全连接层。所有的卷积层都使用3×3的卷积核和2×2的步长，逐渐减

低输入帧的分辨率。全连接层使用Sigmoid作为激活函数，获得输入为真实帧的概率。

身份保持器 为了保证生成帧能够兼容行人重识别任务，STCnet要求生成帧具有正确的身份信息。为此，STCnet引入了一个身份保持器（Identity Guider）。身份保持器以生成器的输出作为输入，输出生成帧的身份分类结果，并使用交叉熵损失约束分类结果的正确性。身份保持器可以是任意已有的行人重识别网络。本方法采用一个基础的行人重识别网络 [43]，使用ResNet50 [85] 作为主干网络，并将分类层的输出修改为训练集的身份个数。

4.2.2 损失函数

STCnet的核心是如何学习生成器，使其能将输入的遮挡帧$\widetilde{\mathbf{x}}$复原为相应的未遮挡帧$\mathbf{x}$。但是对于给定遮挡帧$\widetilde{\mathbf{x}}$，训练集中并没有对应的未遮挡帧$\mathbf{x}$。为了构造这样的训练对，本方法首先将训练集中的未遮挡帧随机擦除一个区域，形成对应的遮挡帧。形式化如下：

$$\widetilde{\mathbf{x}} = (1 - \mathbf{M}) \odot \mathbf{x}, \tag{4.3}$$

其中，$\mathbf{M}$表示一个二值掩码，将擦除区域内的值设为1，其余区域的值设为0。给定遮挡帧$\widetilde{\mathbf{x}}$和对应的未遮挡帧$\mathbf{x}$，一方面，我们希望将$\widetilde{\mathbf{x}}$的遮挡区域复原成真实的身体区域。为了达到目的，本方法使用**重构损失**来保证生成图像$\hat{\mathbf{x}}$修复出被遮挡的身体内容。与此同时，本方法还使用**对抗损失**来保证$\hat{\mathbf{x}}$的视觉真实性。另一方面，理想的遮挡补全应该保持行人的身份不变。为此，本方法引入了**身份预测约束**来保证$\hat{\mathbf{x}}$与$\mathbf{x}$的身份一致。

重构损失 本方法通过重构损失引导空间生成器G_s和时序生成器G_t修复出被遮挡的身体部件内容。具体地，本方法希望G_s和G_t能够学会从$\widetilde{\mathbf{x}}$复原出$\mathbf{x}$。为了降低生成难度，生成器保留了输入帧中未被遮挡的像素，只预测遮挡区域的像素值。因此，有如下目标函数：

$$\begin{aligned} \mathcal{L}_r &= \min_{G_s, G_t} \|\hat{\mathbf{x}}_\mathbf{1} - \mathbf{x}\|_1 + \|\hat{\mathbf{x}} - \mathbf{x}\|_1 \\ \hat{\mathbf{x}}_\mathbf{1} &= \mathbf{M} \odot G_s(\widetilde{\mathbf{x}}) + (1 - \mathbf{M}) \odot \widetilde{\mathbf{x}} \\ \hat{\mathbf{x}} &= \mathbf{M} \odot G_t\left(\hat{\mathbf{x}}_\mathbf{1}, \mathbf{x_p}, \mathbf{x_n}\right) + (1 - \mathbf{M}) \odot \widetilde{\mathbf{x}}, \end{aligned} \tag{4.4}$$

其中，$\hat{\mathbf{x}}_\mathbf{1}$和$\hat{\mathbf{x}}$分别表示空间生成器和时序生成器输出的帧，$\mathbf{x_p}$和$\mathbf{x_n}$分别表示$\mathbf{x}$的前一帧和后一帧，此处使用$\ell_1$损失而不是$\ell_2$，是因为$\ell_1$损失可以更好地抑制模糊

现象 [112]。

对抗损失　本方法引入了生成器和全局判别器D_g之间的对抗学习，旨在保证生成图像$\hat{\mathbf{x}}$的整体真实性。此外，引入了生成器和局部判别器D_l之间的对抗损失，旨在保证遮挡区域生成内容的有效性。本方法沿用了 [119] 的对抗损失，其形式化如下：

$$\begin{aligned}\mathcal{L}_{a_1} &= \min_{G_s,G_t}\max_{D_g}\mathbb{E}_{\mathbf{x}\sim p_{data}(\mathbf{x})}\left[\log D_g(\mathbf{x}) + \log\left(1 - D_g(\hat{\mathbf{x}})\right)\right] \\ \mathcal{L}_{a_2} &= \min_{G_s,G_t}\max_{D_l}\mathbb{E}_{\mathbf{x}\sim p_{data}(\mathbf{x})}\left[\log D_l(\mathbf{M}\odot\mathbf{x}) + \log\left(1 - D_l(\mathbf{M}\odot\hat{\mathbf{x}})\right)\right],\end{aligned} \tag{4.5}$$

其中 $P_{data}(\mathbf{x})$表示真实帧$\mathbf{x}$的分布。

身份预测损失　本方法要求生成图像$\hat{\mathbf{x}}$的身份保持不变。为此，本方法使用一个身份保持器R约束$\hat{\mathbf{x}}$维持行人身份。身份预测损失通过交叉熵实现：

$$\mathcal{L}_c = \min_{G_s,G_t} -\sum_{k=1}^{K}\mathbf{q}_k \log R(\hat{\mathbf{x}})_k, \tag{4.6}$$

其中，K为训练集的行人身份个数，$\mathbf{q}$为$\mathbf{x}$的真实身份标签。

总体损失　最终，STCnet通过上述的重构损失、对抗损失以及身份预测损失联合训练，整体的目标损失为：

$$\mathcal{L} = \mathcal{L}_r + \lambda_1(\mathcal{L}_{a_1} + \mathcal{L}_{a_2}) + \lambda_2\mathcal{L}_c, \tag{4.7}$$

λ_1和λ_2是用来平衡各个损失函数的超参数。

4.3　遮挡鲁棒视频行人重识别框架

通过结合时空补全网络和行人重识别网络，本章提出了一个遮挡鲁棒视频行人重识别框架VRSTC。如图1.6所示，VRSTC可以划分成三个阶段：（1）遮挡区域定位：通过相似性打分机制定位视频帧中发生遮挡的区域；（2）遮挡补全：利用时空补全网络修复遮挡区域的内容；（3）行人重识别：联合补全帧和原始未遮挡帧训练一个行人重识别模型。

遮挡区域定位　真实场景中，同一个遮挡模式一般只出现在一段时间内。比如，静态障碍物的遮挡只出现在行人经过该障碍物的那段时间。因此，对于整个行人序列，同一个遮挡物通常只在连续的几帧内存在，并且遮挡区域会存在剧烈的外观变化。基于这个观察，本章提出一个相似性打分机制，利用遮挡

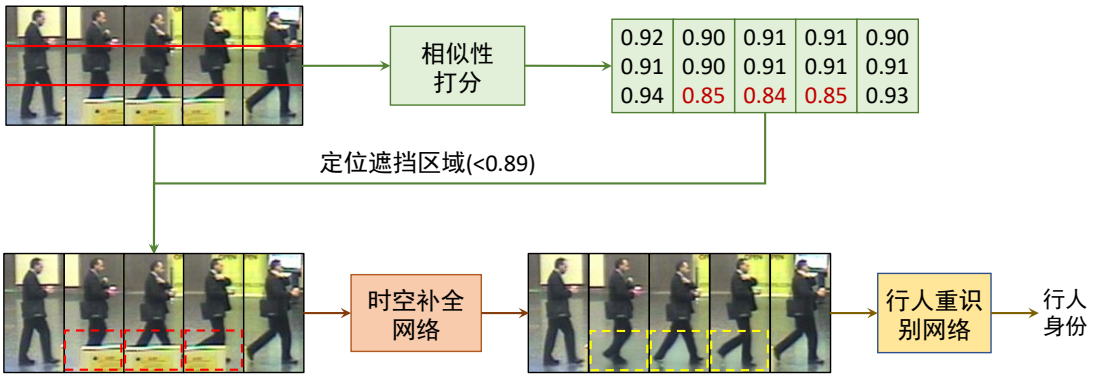


图 4.4 遮挡鲁棒的视频行人重识别框架VRSTC的结构图

区域的外观变化性质对其赋予一个低的权重。具体实现上，给定一个行人序列$I = \{I_t\}_{t=1}^{T}$，其中T为行人序列的帧数。首先，将每一个视频帧垂直划分为三个固定的局部区域$I_t = \{I_t^u, I_t^m, I_t^l\}$，其中$u, m$和$l$分别指示帧的上、中、下部分。然后使用卷积网络提取每个区域的特征表示$\{\mathbf{v_t^k} | k \in \{u, m, l\}$，并通过时序平均池化得到视频级别的区域特征$\overline{\mathbf{v^k}}$：

$$\overline{\mathbf{v^k}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{v_t^k}, \quad k \in \{u, m, l\}. \tag{4.8}$$

最后，根据$\mathbf{v_t^k}$与$\overline{\mathbf{v^k}}$的余弦相似值衡量对应区域发生遮挡的概率。由于遮挡物存在剧烈的表观变化，相似性越低说明对应区域发生遮挡的概率越大：

$$u_t^k = \left\langle \frac{\mathbf{v_t^k}}{||\mathbf{v_t^k}||_2}, \frac{\overline{\mathbf{v^k}}}{||\overline{\mathbf{v^k}}||_2} \right\rangle. \tag{4.9}$$

本章方法将u_t^k小于0.89的局部区域视为被遮挡的区域。在定位到每一帧的遮挡区域之后，VRSTC将包含遮挡区域的序列输入到STCnet中复原出未遮挡帧，构成一个新的数据集。最后，使用这个新的数据集训练一个行人重识别模型。

行人重识别模型 本方法使用一个最基础的视频行人重识别模型 [43]。实现上，使用ResNet50提取每一帧的特征，然后通过时序平均池化整合多帧特征。

4.4 Occluded-DukeMTMC-VideoReID数据集组织与分析

为了促进遮挡行人重识别的研究，本章重新组织了DukeMTMC-VideoReID数据集 [76]，构建了一个大规模基于视频的遮挡行人重识别数据集，Occluded-DukeMTMC-VideoReID。

表 4.1 Occluded-DukeMTMC-Video查询集和候选集上具有一定比例遮挡帧的序列占比

序列中遮挡帧比例 (%)	0	0-40	40-60	60-70	70-80	80-90	90-100
查询集合 (%)	0	1.1	6.1	8.4	13.8	33.2	33.4
候选集合 (%)	31.0	25.6	3.4	3.6	5.9	12.4	18.1

表 4.2 Occluded-DukeMTMC-Video查询集和候选集上具有一定遮挡率的视频帧占比

单帧遮挡率 (%)	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-100
查询集合 (%)	13.5	11.1	12.4	14.8	14.9	15.6	11.1	6.6
候选集合 (%)	13.3	11.2	12.3	14.7	14.4	14.6	11.2	8.3

4.4.1　数据集组织

按照Occluded-DukeMTMC数据集 [22] 的组织过程，本章重新划分DukeMTMC-VideoReID 的测试集，使所有的查询序列都包含遮挡。具体来说，从原始数据集的查询集和候选集中人工选择包含遮挡的序列作为新的查询序列。查询集的构建分为三个步骤：（1）将包含多个行人，以及被静态障碍物（例如树木或汽车）遮挡的视频帧标注为遮挡帧；（2）从原始数据集的查询集和候选集中选出所有包含遮挡帧的序列；（3）从每个选中的序列中随机裁剪一个子序列，同时约束裁剪的子序列包含1/3以上的遮挡帧，将裁剪后的子序列作为Occluded-DukeMTMC-VideoReID的查询序列。

构建完查询集之后，将原始测试集中的剩余序列作为候选集。在构建训练集时，手动删除掉原始训练集的494个序列。这是由于这494个序列包含与测试集完全相同的障碍物，这些训练序列可能导致模型记忆特定的遮挡模式，从而高估了训练模型的泛化性。

4.4.2　数据统计分析

Occluded-DukeMTMC-VideoReID数据集最终包括1,812个行人和4,338个序列。其中训练集覆盖了702个行人和1, 702个序列；查询集包含661个行人，每个行人只有一个序列；候选集包含1, 110个行人和2, 636个序列。该数据集是第一个基于视频的遮挡行人重识别数据集，也是目前最大的遮挡行人重识别数据集。

由于研究者更加关注模型在测试集上的表现，本节给出了Occluded-DukeMTMC-VideoReID测试集的遮挡统计分析。（1）本节统计了查询集和候选集中遮挡序列

的比例。根据章节4.4.1所述，查询集100%的序列都包含遮挡。此外，本章统计出候选集70%以上的序列都存在遮挡。（2）本节标注了查询集和候选集中每个序列的遮挡帧比例。如表4.1所示，60%/30%以上的查询/候选序列包含80% – 100%的遮挡帧。统计结果表明了Occluded-DukeMTMC-VideoReID测试集包含严重的遮挡，因此能够评估模型在遮挡场景的性能。（3）本节标注了查询集和候选集中每个视频帧的遮挡分数$\{0, 1, 2, \ldots, 10\}$，其中0表示帧不存在遮挡，1表示帧的0 – 10%区域存在遮挡，依此类推。如表4.2所示，65%以上的帧的遮挡区域小于50%，表明部分遮挡更为常见。

4.5 实验与分析

本节分别在MARS [43]、DukeMTMC-VideoReID [76]和本章组织的Occluded-DukeMTMC-VideoReID三个视频数据集上对本章方法VRSTC进行评测。本节接下来分别介绍了实验细节、消融实验、参数分析、与其他方法的对比实验和可视化分析。

4.5.1 实现细节

VRSTC的训练过程分为以下三个阶段。第一阶段：预训练身份保持器。本章使用交叉熵损失训练一个以ResNet50为主干网络的行人重识别网络，将其作为时空补全网络的身份保持器。在训练阶段，本章采用Adam [95] 优化器进行网络优化，共训练150个回合。初始学习率设为3.0×10^{-4}，每训练50个回合后下降10倍。训练使用的批量大小为32，每个输入序列通过在原始视频中随机采样4帧得到。本章只采用水平翻转作为数据增广。

第二阶段：训练时空补全网络。首先，以预训练的行人重识别网络作为特征提取器，使用相似性打分机制生成每个局部区域的分数，并将分数小于0.89的区域视为发生遮挡的区域；然后，将训练集中的未遮挡帧随机擦除一个局部区域作为遮挡帧，人工构造训练时空补全网络所需的数据。时空补全网络以预训练的行人重识别网络作为身份保持器，使用Adam优化器进行训练，学习率设置为0.0001。输入图像大小为128×64。超参数λ_1和λ_2分别设置为0.001和0.1。

第三阶段：使用补全帧训练行人重识别网络。本阶段将原始数据集中的遮挡帧替换为时空补全网络的生成帧，构成一个新的数据集。行人重识别网络则根据这个新的数据集进行训练和测试。本阶段在ResNet50主干网络中嵌

(a) 自行车遮挡　　(b) 行人遮挡

图 4.5 MARS数据集上遮挡视频示例

入NonLocal [92] 模块作为行人重识别网络。其他的训练设置与第一阶段一致。

测试阶段　首先将包含遮挡的测试序列送入到时空补全网络中修复遮挡帧，形成一个补全后的视频序列；然后将这个补全的测试序列等分为多个含有4帧的视频片段，利用训练完成的行人重识别网络提取每个视频片段的特征；最后将所有视频片段的特征取平均得到测试序列的视频特征。

对于每个查询视频，计算它与所有候选视频的特征相似性。然后根据计算出的相似性对所有的候选视频进行排序，选出最相似的候选视频，达到目标检索的目的。

4.5.2　时空补全网络的消融实验

本节在MARS数据集上采取一系列消融实验，以验证STCnet各个模块的有效性。 MARS数据集在清华校园内采集，存在静态障碍物以及其他行人遮挡的情况。如图4.5(a)所示，校园中放置大量的自行车，因而很多行人存在被自行车遮挡的样例。此外，如图4.5(b)所示，上下课时间同学们会比较聚集，校园场景也存在行人遮挡情况。

本节首先引入一个基准模型Baseline，采用与VRSTC相同的行人重识别模型，但使用原始数据进行训练和测试。为了公平比较，Baseline使用与本章方法相同的训练设置。接下来，本节对STCnet的各个组件进行有效性分析。实验结果如表4.3所示。

空间生成器的有效性　Spa将空间生成器作为补全网络，并且只使用空间重构损失训练。相比于Baseline，Spa在top-1和mAP上分别提升了1.1%和0.9%。实验表明空间生成器能够在一定程度上恢复遮挡区域的判别内容，提升行人重识别的性能。

时序生成器的有效性　Spa+Tem将空间生成器和时序生成器的组合作为补全网络，使用空间和时序重构损失联合训练。从表4.3可以观察到，引入时序生

表 4.3 时空补全网络STCnet的消融分析

方法	MARS	
	mAP (%)	top-1 (%)
基础行人重识别模型（Baseline)	79.9	86.1
空间生成器作为补全网络 (Spa)	81.0	87.0
时序生成器的变体：自动编码器（Spa+AE）	80.8	87.0
时序生成器的变体：时序自动编码器（Spa+TAE）	81.0	87.3
空间+时序生成器作为补全网络 (Spa+Tem)	81.6	87.8
加入局部判别器（Spa+Tem+LD）	81.7	87.9
加入局部和全局判别器（Spa+Tem+LD+GD）	81.9	87.9
STCnet	**82.3**	**88.5**

成器能够带来一致的性能提升。相比于Spa，Spa+Tem提升了0.8%top-1。这是由于时序生成器关注到相邻帧中与遮挡部件高度相关的内容，使生成帧的语义与视频保持一致。因而行人重识别网络能够更好考察补全后序列的时序信息，生成一个更具判别性的行人特征。

为了证明Spa+Tem性能的提升并不仅仅是来自网络深度的增加，本节引入了时序生成器的两个变体：“自动编码器（Autoencoder，AE）”和“时序自动编码器（Temporal Autoencoder，TAE）”。AE只以空间生成器的输出作为输入，没有考虑视频的时序信息。AE是一个标准的编码器-解码器，采用与时序生成器相同的编码器和解码器结构，因而具有与时序生成器相同的参数量。如表4.3所示，AE不能带来性能的提升，说明Spa+Tem的性能提升并不是由于网络深度的增加。TAE在时序生成器中去掉了时序注意力层。如表4.3所示，Spa+TAE的性能低于Spa+Tem，验证了时序注意力层的有效性。

判别器的有效性 Spa+Tem+LD联合生成器和局部判别器作为补全网络，Spa+Tem+LD+GD联合生成器，局部判别器和全局判别器作为补全网络。这两个网络都使用重构损失和对抗损失进行训练。从表4.3可以观察到，判别器仅能带来微小的性能提升。这是因为判别器的目标是生成视觉真实的图像，不会引入太多利于重识别的判别线索。

身份保持器的有效性 STCnet结合生成器，判别器和身份保持器，并且使

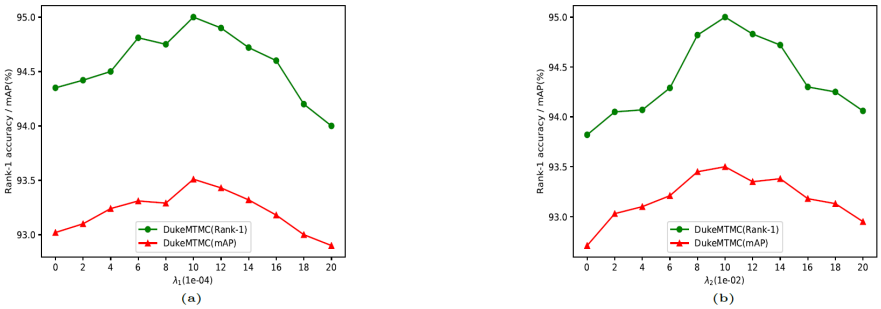


图 4.6 DukeMTMC-VideoReID数据集上超参数λ_1和λ_2的影响

表 4.4 本章方法与当前视频行人重识别方法在Occluded-DukeMTMC-VideoReID的性能对比

	方法	Occluded-DukeMTMC-VideoReID			
		mAP (%)	top-1(%)	top-5 (%)	top-10 (%)
集合建模	TriNet [20]	64.1	63.1	82.6	87.4
	STAN [46]	69.4	69.4	88.1	91.7
	QAN [44]	74.8	75.1	90.6	93.4
	RQEN [45]	74.9	73.5	90.5	94.1
时空建模	RCN [48]	62.4	60.9	83.3	88.1
	Snippet [61]	72.5	74.1	89.2	92.1
	IANet [68]	81.2	81.9	93.9	95.6
	TCLNet [120]	82.1	82.0	93.5	96.2
	VRSTC（本章方法）	**82.7**	**84.0**	**94.4**	**97.1**

用重构损失、对抗损失和身份预测损失联合训练。如表4.3所示，引入身份保持器在MARS数据集上带来了0.6%top-1和0.4%mAP提升。实验表明了身份保持器有利于保留与行人身份关联的视觉线索，使得生成帧更容易识别。

4.5.3　超参数分析

本节测试了超参数的影响。超参数λ_1和λ_2分别用来控制对抗损失和身份约束损失的影响。图4.6测试了λ_1和λ_2在DukeMTMC-VideoReID数据集上对性能的影响。可以观察到，当$\lambda_1 = 0.001$和$\lambda_2 = 0.1$时，模型取得了最好的重识别性能。值得注意的是，过大的λ_1和λ_2都会导致性能的大幅降低。这主要是因为当对抗损失或者身份约束损失占有主导因素时，时空补全网络难以收敛，导致重识别性能严重降低。

表 4.5 本章方法与当前视频行人重识别方法在MARS和DukeMTMC-Video的性能对比

	方法	MARS		DukeMTMC-Video	
		mAP (%)	top-1 (%)	mAP (%)	top-1 (%)
集合建模	QAN [44]	51.7	73.7	-	-
	STAN [46]	65.8	82.3	-	-
	EUG [76]	67.4	80.8	78.3	83.6
	RQEN [45]	71.7	77.8	-	-
	TAFD [47]	78.2	87.0	-	-
时空建模	Snipped [61]	69.4	81.2	-	-
	M3D [52]	74.1	84.4	-	-
	Snippet+OF [61]	76.1	86.3	-	-
	GLTP [94]	78.5	87.0	93.7	96.3
	IANet [68]	85.0	**90.2**	96.1	**96.9**
	TCLNet [120]	**85.1**	89.8	**96.2**	**96.9**
	VRSTC （本章方法）	82.3	88.5	93.8	95.0

4.5.4 与当前先进方法的对比

本节将本章方法VRSTC与当前视频行人重识别方法进行比较。如表4.4所示，VRSTC在Occluded-DukeMTMC-VideoReID数据集上取得了最高的性能水平。可以观察到：（1）相比于基于循环网络的时空建模方法Snippet [61]，基于时间注意力的方法QAN[44]和RQEN[45]取得了更高的性能。实验结果表明时间注意力机制可以在一定程度上提高模型对遮挡的鲁棒性。（2）相比于基于时间注意力的方法RQEN [45]，VRSTC提升了7.8%mAP和10.5%top-1。性能的大幅提升验证了相比于已有的抑制遮挡帧策略，时空补全策略可以更好地解决视频行人重识别的严重遮挡问题。（3）相比于第二章的IANet [68] 方法以及第三章的TCLNet [120] 方法，VRSTC取得了更高的准确率，验证了时空补全策略在严重遮挡场景上的优越性。

与前两章方法的对比分析 如表4.5所示，在MARS和DukeMTMC-VideoReID数据集上，VRSTC取得了较高的性能水平，但是性能低于上两章的IANet和TCLNet方法。这是由于这两个数据集只有不到10%的序列存在遮挡，因而限制了图像时

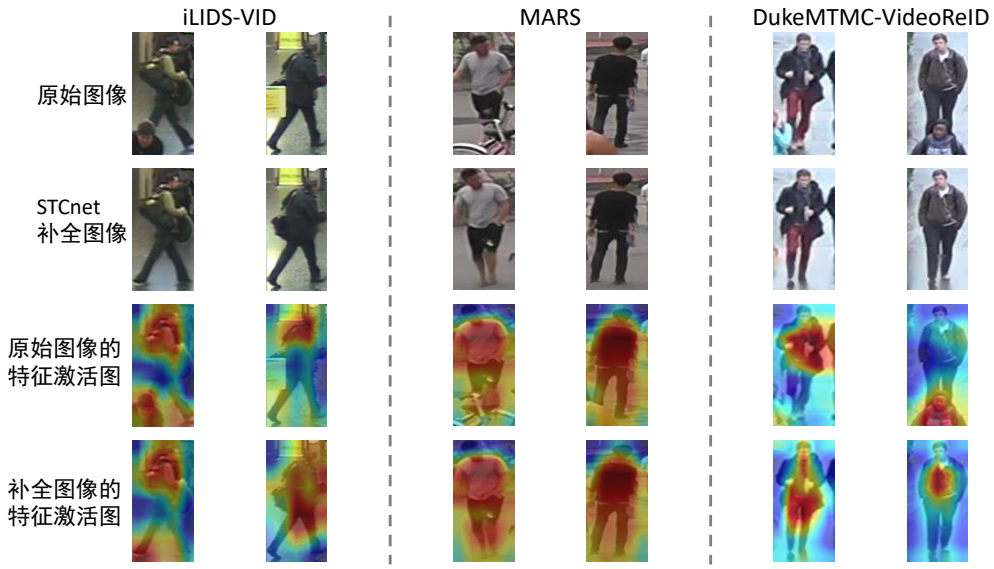


图 4.7 STCnet补全图像和特征激活图的可视化

空补全的效果。此外，通过对比这三个方法在DukeMTMC-VideoReID和Occluded-DukeMTMC-VideoReID数据集上的性能可以发现，相比于上两章的IANet和TCLNet方法，VRSTC在严重遮挡数据集上的性能下降幅度更小。实验结果表明了VRSTC具有更好的遮挡鲁棒性。

4.5.5　可视化分析

图4.7展示了STCnet对遮挡帧的修复结果，以及原始帧和补全帧的特征激活图。可以观察到STCnet在一定程度上可靠复原了遮挡区域的像素，并且改善了目标行人的特征表示。（1）当一个行人被其他行人的某个身体部位遮挡时，目标行人的特征通常会掺杂其他行人的视觉外貌。如图4.7的第六列所示，干扰行人的头部被重识别模型激活，损害了目标行人的特征表示。（2）当行人被广告牌、自行车等静态物体遮挡时，行人的特征会丢失大量身体信息。如图4.7的第二列所示，行人特征只关注到头肩区域，丢失了下半身信息。（3）从图4.7的最后一行可以观察到，当STCnet修复了遮挡区域的像素，行人重识别模型有机会考虑更多的身体区域，从中发现新的判别线索帮助识别目标行人。

4.6　本章小结

本章探索了面向遮挡场景的视频行人重识别问题，根据行人序列的空间结构和时间关联设计了一个时空补全网络STCnet。STCnet将重构学习、对抗学习

和身份预测约束巧妙地联合到一起，进而形成一个有效的遮挡区域修复模型。基于时空补全网络，本章提出了一个遮挡鲁棒视频行人重识别框架。该框架可以与任意先进的视频行人重识别方法相结合，具有一定的通用性。本章的工作表明对于解决行人重识别遮挡问题，在以部分遮挡为主导的条件下，时空补全遮挡区域相比于常用的丢弃遮挡帧策略更为有效。

本章提出的方法也有一定的局限性，主要体现在两个方面：1）STCnet主要目的在于生成视觉真实的图像，因而更加关注生成帧的视觉质量而非判别性，导致补全操作并没有充分提升重识别精度；2）行人重识别需要一个昂贵的图像补全处理，导致整个重识别系统的计算开销过大，不利于真实场景的部署。下一章工作探讨了如何避免昂贵的图像生成去实现遮挡区域的修复，提出从特征补全的角度改进本章方法。

第5章　面向遮挡场景的特征时空补全方法

上一章提出的方法虽然初步缓解了行人重识别的遮挡问题，然而存在图像补全与行人重识别的目标不完全一致问题。本章提出了一个特征时空补全机制，通过设计合理的网络模块利用时空信息预测遮挡区域的特征。特征补全方法避免了昂贵的图像生成步骤，进一步提升了遮挡场景下行人重识别的鲁棒性和效率。

5.1　引言

本章将继续研究遮挡场景下的视频行人重识别任务。上一章设计了一个图像时空补全网络去恢复遮挡区域的像素，在一定程度上解决了视频行人重识别的遮挡问题。但是上章方法存在图像补全与行人重识别目标不完全一致的问题：图像补全的主要目标是保证生成帧的视觉真实性；行人重识别的目标是提高遮挡区域的特征判别力。由于视觉真实的图像并不一定最具判别力，图像补全与行人重识别模型没有完全兼容，因而只能带来有限的性能提升。针对这个问题，本章提出一个特征层面的时空补全学习方法。

相比于图像补全 [117, 121]，特征补全具有如下两个优势。第一，特征补全能够无缝嵌入到行人重识别模型中，实现端到端优化。如图5.1(a)所示，图像补全网络通常堆叠多个卷积层，参数量巨大。因此图像补全网络一般单独训练，作为行人重识别模型的数据前处理步骤。在这种设置下，行人重识别模型不能给图像补全任务提供直接的反馈信息，导致图像补全更加专注于生成视觉真实的图像而非提升重识别性能。相比之下，轻量级的特征补全能够嵌入到行人重识别模型中，使得行人重识别的反馈信号可以直接监督特征补全操作，充分提升行人重识别模型的性能。

第二，特征补全能够捕捉长期的空间上下文和时间上下文关系。根据上一章的介绍，图像补全采用编码器-解码器结构，其中编码器和解码器分别由多个卷积层堆叠而成。一方面，由于卷积运算的局部性，只有堆叠多个卷积层才能捕捉到远距离的空间上下文。如图5.1(a)所示，可视的“黄色”像素的信息只能通过堆叠多个卷积层才能传播到遮挡的“红色”像素。然而重复卷积操作不仅

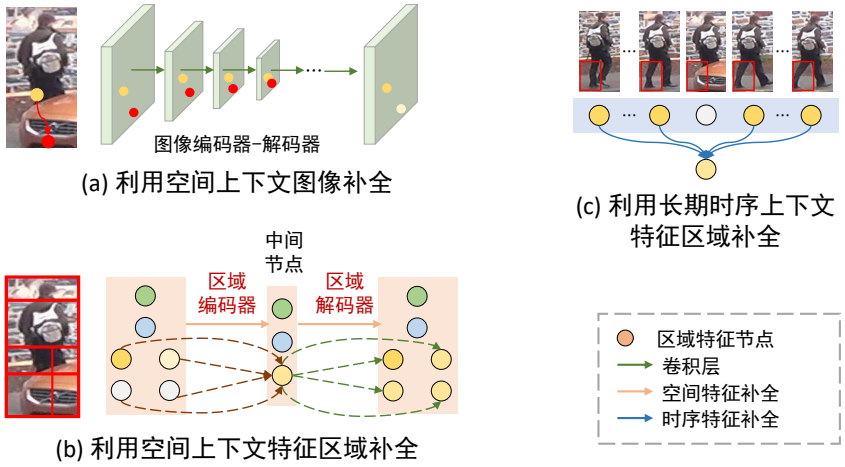


图 5.1 本章方法的动机：相比于图像补全，特征补全能够捕捉长期的空间和时间上下文

计算效率低下而且会造成优化困难 [92]。相反，由于特征图具有一定的语义信息，相同语义的特征单元能够整合形成一个表示特定身体部件的区域特征。如图5.1(b)所示，卷积特征图可以划分为少数几个区域特征，利用区域特征更容易在远距离的身体部件间传播信息。因此，特征补全可以更有效地捕捉远距离的空间上下文；另一方面，由于计算资源的限制，图像补全通常只考虑相邻两帧信息。如图5.1(c)所示，特征补全只需预测少数几个区域特征，具有非常少的计算开销，因此能够利用更多的相邻帧信息。综合上述分析，特征补全可以捕捉更长期的时空上下文信息，从而更有利于修复遮挡区域的特征。

基于上述的分析，本章认为图像层面的遮挡修复对行人重识别任务来说是次优且不必要的，而行人重识别需求的是遮挡区域的特征被正确修复。因此，本章提出一个区域特征补全方法（Region Feature Completion，RFC），旨在直接修复遮挡区域的特征表示。首先，本章设计了一个空间区域特征补全模块（Spatial Region Feature Completion，SRFC），由区域编码器和区域解码器组成。如图5.1(b)所示，区域编码器学习将遮挡区域和相关可视区域聚合到一个中间节点，区域解码器则从这个中间节点中修复遮挡区域的特征。以中间节点作为媒介，关联的可视区域信息能够传播到遮挡区域中，帮助修复遮挡区域的特征表示。其次，本章设计了一个时序区域特征补全模块（Temporal Region Feature Completion，TRFC），利用长期的时间上下文对SRFC生成的特征进行精细化。

不同于昂贵的图像补全网络，本章提出的RFC模块只增加了极少的计算开销，并能够嵌入到任意已有行人重识别模型中，提高模型对遮挡的鲁棒性。实

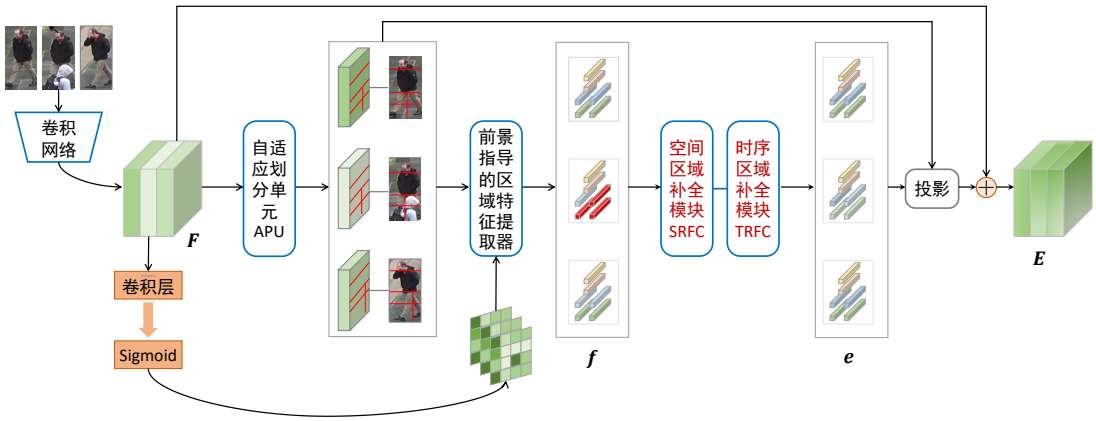


图 5.2 区域特征补全模块（Region Feature Completion，RFC）结构图

验表明，相比于上一章的图像补全方法，本章设计的特征补全方法可以更好地解决行人重识别的遮挡问题。

本章接下来安排如下：章节5.2介绍本章所提出的区域特征补全模块RFC；章节5.3介绍基于RFC的行人重识别网络；章节5.4给出本章方法在公开数据集上的评测结果；章节5.5将对本章内容进行小结。

5.2　区域特征补全模块

这一节详细介绍本章提出的RFC方法。如图5.2所示，RFC主要包含四个组件：自适应划分单元、前景指导的区域特征提取器、空间区域特征补全模块和时序区域特征补全模块。以下内容将阐述RFC各个部分的设计原理并介绍用以指导模型学习的目标函数。

5.2.1　自适应划分单元

上一章的方法采用均匀划分策略，即均匀划分行人图像，然后分析每个划分的局部区域是否包含遮挡。但是，由于行人的姿态和尺度变化，均匀划分会导致不同帧中相同位置的局部区域对应不同的语义。针对这个问题，本章提出了一个自适应划分单元（Adaptive Partition Unit，APU），根据人体的拓扑结构将输入的特征图自适应划分为多个局部区域，使每一个局部区域对应一个指定的身体部件。由于遮挡通常发生在行人的下半身，本方法对行人的下半身进行更精细划分。具体实现上，在行人帧上定义N个区域，分别对应行人的头部，上身，左腿上部分，左腿下部分，右腿上部分和右腿下部分。给定输入特征图$F_t \in \mathbb{R}^{D\times H\times W}$，其中$t$表示输入序列的第$t$帧，$D$、$H$、$W$分别表示特征图的通道

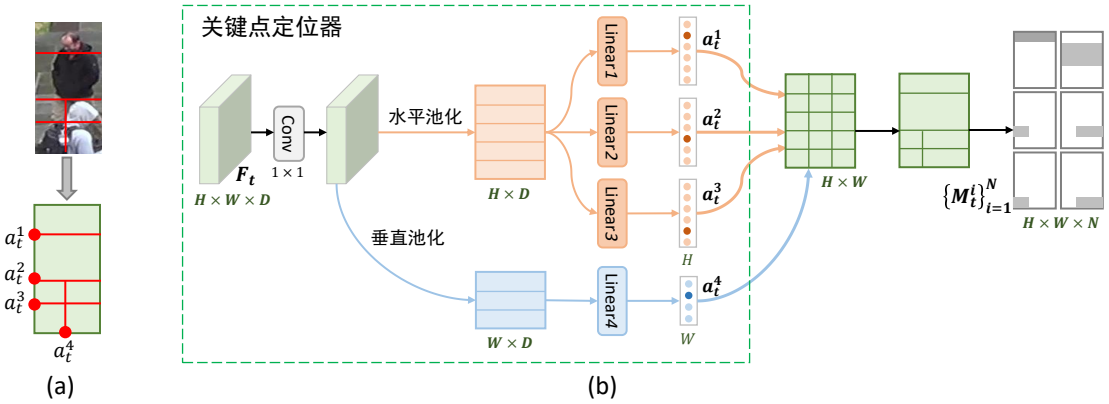


图 5.3 自适应划分单元（Adaptive Partition Unit，APU）结构图

数，空间长度和宽度。APU首先使用一个关键点定位器定位预定义区域在F_t中的位置，然后为每个区域生成对应的二值掩码。

如图5.3(a)所示，对于一个行人图像，四个关键点，$\{(a_t^1,0),\ (a_t^2,0),\ (a_t^3,0),\ (0,a_t^4)\}$，便足够划分出预定义的区域。为此，本章设计了一个关键点定位器，旨在预测这四个关键点的位置。如图5.3(b)所示，对于列关键点$\{a_t^i\}_{i=1}^3$，首先沿着水平方向对F_t执行池化操作，得到F_t的列描述符。列描述符包含了行人垂直方向的信息，因此能够有效预测出垂直方向的关键点。本方法将列描述符输入到三个线性层，并使用Softmax函数生成列关键点对应的概率分布$\{p_t^i\}_{i=1}^3$。最后取$\{p_t^i|i\ =\ 1,2,3\}$最大值对应的索引位置得到列关键点$\{a_t^i|i\ =\ 1,2,3\}$：$a_t^i\ =\ \arg\max\left(p_t^i\right)$。类似地，对于行关键点$a_t^4$，首先沿着垂直方向对$F_t$执行池化操作获得$F_t$的行描述符，然后将行描述符输入到一个线性层中得到概率分布p_t^4，最后取$a_t^4\ =\ \arg\max\left(p_t^4\right)$。根据关键点定位器生成的关键点，每个特征单元$(F_t)_{hw}$能够被分类到预定义的区域$\{R_t^i\}_{i=1}^N$中。最后，对于每个区域$R_t^i$，通过将区域内的值置为1，生成该区域的二值掩码$M_t^i\in\mathbb{R}^{H\times W}$。

5.2.2 前景指导的区域特征提取器

前景指导的区域特征提取器用于提取APU划分区域的特征。如图5.2所示，F_t经过一个以$1\times1\times D\times1$为卷积核，Sigmoid为激活函数的卷积层生成一个前景图$I_t\in\mathbb{R}^{H\times W}$。本章使用行人分割模型 [30] 指导$I_t$的生成。在行人分割模型的指导下，$I_t$可以学会对行人身体区域分配相对较大的值，对遮挡物和背景区域分配

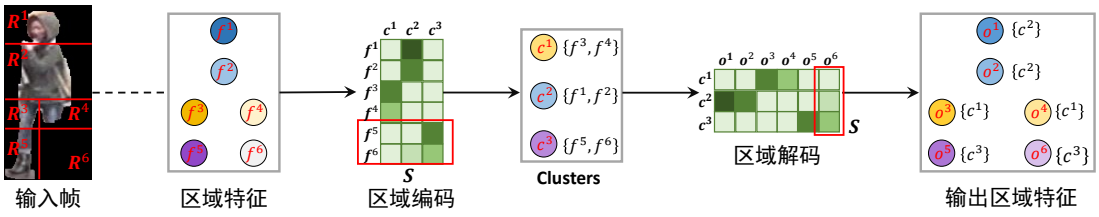


图 5.4 空间特征补全模块的编码和解码过程

较小的值。因此利用I_t可以获得一个判别的区域特征$f_t \in \mathbb{R}^{N\times D}$，

$$f_t^i = (f_t)_i = \frac{M_t^i \odot I_t}{\left\| M_t^i \odot I_t \right\|_1} * F_t, \tag{5.1}$$

其中$f_t^i \in \mathbb{R}^D$表示第t帧的第i个区域的特征向量，$\odot$表示逐元素乘法操作，$*$表示加权求和运算。

5.2.3　空间区域特征补全模块

尽管前景指导的提取器能够抑制遮挡物的特征，减轻遮挡物对行人特征的破坏，但是遮挡仍然会造成目标行人的信息损失。针对这个问题，本章设计了一个空间区域特征补全模块SRFC，目的是利用人体拓扑结构修复遮挡区域的特征。SRFC独立处理输入序列的每一帧，为了简单起见，本节省略了所有符号的时间下标t。

如图5.4所示，SRFC由两个子网络组成：一个区域编码器和一个区域解码器。区域编码器将所有区域特征$\{f^i\}_{i=1}^N$聚合成若干个簇（cluster）$\{c^k\}_{k=1}^K$。区域编码过程通过一个分配矩阵$S \in \mathbb{R}^{N\times K}$实现，其中$S_{ik}$表示将第$i$个区域分配给第$k$个簇的概率，$S_i \in \mathbb{R}^K$表示第$i$个区域的分配向量。区域编码的目标是如果不同区域具有相似的外观或相近的位置，则为这些区域输出相似的分配向量。通过这种方式，区域编码器能够将关联区域分配到同一个簇。此时，每个簇聚集了高度相关的区域特征，能够代表一个较大的身体区域。然后，区域解码器根据分配矩阵将簇$\{c^k\}_{k=1}^K$分布到输出区域特征$\{o^i\}_{i=1}^N$。通过这种方式，区域解码器能够利用关联的簇修复遮挡区域的特征表示。如图5.4所示，区域编码器依靠位置信息将R^6和相近的R^5分配给同一个簇c^3，随后区域解码器基于c^3预测R^6的特征。以c^3作为中介，R^5的信息能够传播到R^6中，帮助修复R^6的特征表示。接下来，本节将具体阐述区域编码器和区域解码器的结构并介绍用以指导其学习的目标函数。

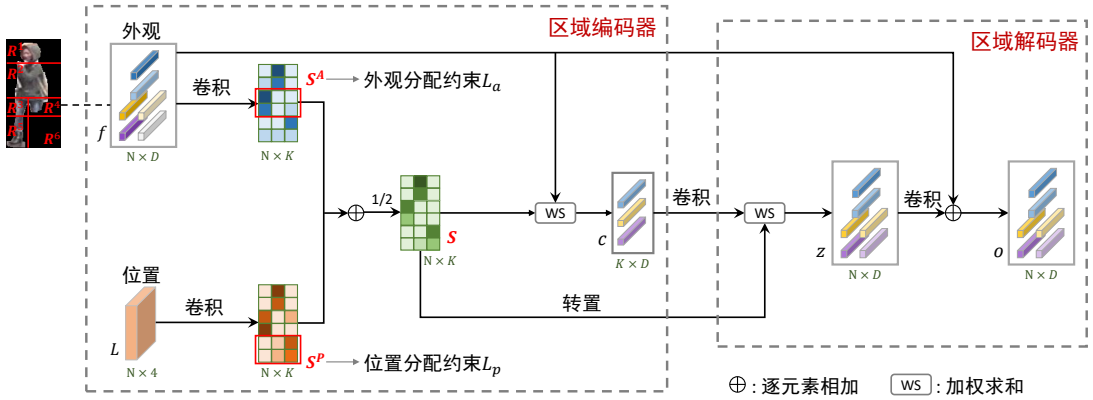


图 5.5 空间特征补全模块（Spatial Region Feature Completion，SRFC）结构图

5.2.3.1 区域编码器

如图5.5所示，区域编码器根据区域的外观信息和位置信息学习一个分配矩阵，并根据分配矩阵将f聚合到若干个簇$\{c^k\}_{k=1}^K$。

外观信息 在区域编码器内部，输入区域特征f经过一层卷积操作学习一个外观分配矩阵$S^A \in \mathbb{R}^{N\times K}$，

$$S^A = \text{Softmax}\left(W^a * f\right), \tag{5.2}$$

其中$*$表示卷积操作，Softmax函数用以保证S_{ik}^A可以代表区域R^i分配到簇c^k的概率。S^A旨在将具有相似外观的区域分配给同一个簇。换言之，如果R^i和R^j的特征具有较高的外观相似性，则对应的分配向量S_i^A和S_j^A也应该高度相似。为此，本章引入了一个外观分配正则项，约束外观分配向量的相似度与对应区域特征的相似度一致：

$$L_a = \sum_i^N \sum_j^N \left\| sim\left(S_i^A, S_j^A\right) - sim\left(f^i, f^j\right) \right\|_1, \tag{5.3}$$

其中sim表示余弦相似度。S^A的主要用处是帮助补全发生部分遮挡区域的特征。如图5.5所示，R^4的一部分发生遮挡，并且剩余可视部分与R^3的外观相似。在L_a约束下，S^A能够将R^4和R^3分配到一个簇。通过这种方式，R^3的信息能够传播到R^4中，帮助修复R^4的特征表示。

位置信息 为了修复完全遮挡的区域，区域编码器利用位置信息关联相近的区域。首先，将R^i的位置信息编码为(x_i, y_i, h_i, w_i)，其中(x_i, y_i)表示区域中心的空间坐标，h_i和w_i表示区域的高度和宽度；然后，组合输入帧N个区域的位置信息，得到一个位置编码$L \in \mathbb{R}^{N\times 4}$；最后，区域编码器根据$L$生成一个位置分配矩

阵$S^P \in \mathbb{R}^{N\times K}$：

$$S^P = \text{Softmax}\left(W_2^p * \max\left(0, W_1^p * L\right)\right), \tag{5.4}$$

其中$W_1^p \in \mathbb{R}^{1\times1\times4\times D}$和$W_2^p \in \mathbb{R}^{1\times1\times D\times K}$为两个可学习的卷积核。 S^P的目标是将相近的区域分配到同一个簇。换言之，如果R^i和R^j的位置相近，对应的位置分配向量S_i^P和S_j^P也应该高度相似。为了实现上述目标，首先给出任意两个区域位置相似度的定义：

$$ps(R^i, R^j) = 1 - 2\sqrt{\left(\frac{y_i - y_j}{H}\right)^2 + \left(\frac{x_i - x_j}{W}\right)^2}. \tag{5.5}$$

然后引入一个位置分配正则项L_p，约束位置分配向量的余弦相似度与对应区域的位置相似度一致：

$$L_p = \sum_i^N \sum_j^N \left\| sim\left(S_i^P, S_j^P\right) - ps\left(R^i, R^j\right)\right\|_1. \tag{5.6}$$

在L_p的约束下，区域编码器倾向于为位置相近的区域生成一致的位置分配向量，从而将相近的区域分配到一个簇。 S^P主要用于补全完全遮挡的区域。如图5.5所示，R^6被完全遮挡以至没有任何外观线索，此时区域编码器可以依靠位置信息。在L_p的约束下，S^p能够将R^6和最邻近的R^5映射到一个簇中。通过这种方式，R^5的信息可以传播到R^6，帮助修复R^6的特征表示。

编码矩阵　分配矩阵S整合了外观信息和位置信息，定义为$S = (S^A + S^P)/2$。如图5.5所示，区域编码器根据S生成多个簇$\{c^k\}_{k=1}^K$。实现上，首先对S沿着列归一化得到编码矩阵$A \in \mathbb{R}^{N\times K}$，然后以$A$的第$k$列为权重对所有区域特征加权求和，生成第$k$个簇的特征表示。整个过程形式化如下：

$$c^k = \sum_{i=1}^N A_{ik} f^i, \quad \text{其中} \quad A_{ik} = \frac{S_{ik}}{\sum_{j=1}^N S_{jk}}. \tag{5.7}$$

从等式5.7可以推断出，每个簇聚集了最相关区域的信息。以这些簇作为中介，SRFC能够建立遮挡区域与远距离的可视区域之间的关联，从而帮助补全遮挡区域的特征。

5.2.3.2　区域解码器

区域解码器是区域编码器的逆运算，其目标是根据分配矩阵S将簇分布到每个区域中。如图5.5所示，首先对S进行转置得到解码矩阵$B \in \mathbb{R}^{K\times N}$，然后根据

解码矩阵将簇转换为补全的区域特征 $z \in \mathbb{R}^{N\times D}$，

$$z^i = \sum_{k=1}^{K} B_{ki} c^k, \quad \text{其中} \quad B_{ki} = S_{ik}, \tag{5.8}$$

由于c^k聚集了关联可视区域的特征，因此遮挡区域能够感知到相关的身体信息修复其特征表示。为了简化补全任务，SRFC采用残差学习机制，定义为$o = f + z$。给定一个输入序列，SRFC对每一帧执行相同操作，生成序列$\{o_1, \ldots, o_T\}$。

5.2.4 时序区域特征补全模块

TRFC模块使用长期的时序线索对SRFC输出结果进行精细化。TRFC基于查询-内存注意力机制，将正在处理的区域特征作为查询向量，其余帧对应的区域特征作为内存。它的核心思想是查询-内存机制会将来自其它帧的时间上下文添加到查询向量中，从而帮助查询向量的特征修复。

TRFC以o_t作为查询向量，计算查询向量与内存中的每一项的点乘相似度[92]，得到两者的匹配概率。然后以匹配概率为权重对内存所有项进行加权求和，获得一个时序上下文向量v_t。整个过程形式化如下：

$$\begin{aligned} \alpha_k &= \frac{\exp\langle o_t, o_k\rangle}{\sum_{l\neq t}\exp\langle o_t, o_l\rangle}, \\ v_t &= \sum_{k\neq t} \alpha_k o_k. \end{aligned} \tag{5.9}$$

不同于一般的查询-内存注意力机制，TRFC使用一个门控单元控制传递信息的比例。在每一个门控单元内部，o_t经过一层卷积操作学习一个门控权重g_t，然后依据g_t对查询向量和时间上下文进行加权求和：

$$\begin{aligned} g_t &= \text{Sigmoid}(W o_t + b), \\ e_t &= g_t \odot o_t + (1 - g_t) \odot v_t. \end{aligned} \tag{5.10}$$

给定SRFC的输出结果$\{o_1, \ldots, o_T\}$，TRFC对每一帧执行相同操作，得到一个输出序列$\{e_1, \ldots, e_T\}$。

为了使上述构建的模块与卷积架构兼容，本方法复用APU生成的区域掩码M_t将e_t投影到原始特征空间。实现上，首先将M_t调整为$N \times HW$大小，然后执行M_t的转置与e_t的矩阵相乘操作，再将生成结果$\overline{F_t}$调整为原始特征大小。最后，本章采用残差学习方式获得每一帧的输出特征E_t，形式化为$E_t = \overline{F_t} + F_t$。

5.3　基于区域特征补全的行人重识别网络

本节介绍了基于区域特征补全的行人重识别网络，区域特征补全网络（Region Feature Completion network, RFCnet）。接下来分别对RFCnet的网络结构和损失函数予以介绍。

5.3.1　网络结构

对于行人重识别任务，本章使用ResNet50 [85] 作为主干网络，使用时序平均池化操作融合多帧特征得到一个视频特征表示。RFC模块维持了输入特征图的尺度，能够嵌入到ResNet50的任意层。考虑到计算复杂度，本章只将RFC放置到特征图发生下采样的瓶颈层。这种放置方式有以下两个优势。一方面，已有遮挡行人重识别工作 [22, 32] 只在网络的最后一层处理遮挡，而RFC模块可以添加在网络的早期阶段，可以从底层减轻遮挡的干扰；另一方面，通过在网络不同阶段放置多个RFC模块，RFCnet能够逐级修复遮挡部件的特征表示，进一步提高模型处理遮挡的能力。

5.3.2　目标函数

训练RFCnet的目标函数包括：（1）行人身份监督的交叉熵损失L_{ce}和困难样本三元组损失L_{tri}；（2）指导SRFC学习的外观分配正则项L_a（等式5.3）和位置分配正则项L_p（等式5.6）；（3）用以保证APU关键点正确预测的关键点损失L_k；（4）用以保证前景图正确生成的前景图损失L_f。其中交叉熵损失和三元组损失的细节见章节2.2.3.1。

关键点损失L_k　APU的目标是预测关键点的位置。本章利用姿态估计模型 [30] 获得关键点的真实位置$\{l_t^i\}_{i=1}^4$，并将$\{l_t^i\}_{i=1}^4$作为APU输出概率$\{p_t^i\}_{i=1}^4$的标签。关键点损失定义如下：

$$L_k = \frac{1}{4T}\sum_{t=1}^{T}\sum_{i=1}^{4} CE\left(p_t^i, l_t^i\right), \tag{5.11}$$

其中CE表示交叉熵损失。

前景图损失L_f　前景图的目标是区分行人图像的背景/遮挡部分和身体部分。本章将行人分割模型 [30] 获得的行人掩码G_t作为前景图I_t的标签，指导前景图的生成。前景图损失通过二值交叉熵实现，定义如下：

$$L_f = -\frac{1}{T}\sum_{t=1}^{T}\left[G_t \log(I_t) + (1 - G_t)\log(1 - I_t)\right]. \tag{5.12}$$

综上，RFCnet的整体目标函数为：

$$L_{all} = (L_{ce} + L_{tri}) + \lambda_1 L_k + \lambda_2 L_f + \lambda_3 (L_a + L_p). \tag{5.13}$$

λ_1, λ_2和λ_3是用来平衡各个损失的超参数。

5.4 实验分析

本节分别在MARS [43] 、DukeMTMC-VideoReID [76] 和Occluded-DukeMTMC-VideoReID三个视频数据集上对本章提出的RFCnet进行评测。为了进一步验证本章方法的泛化性，本章还在一个图像遮挡行人重识别数据集Occluded-DukeMTMC [22] 上进行测试。其中Occluded-DukeMTMC-VideoReID和Occluded-DukeMTMC数据集存在严重的遮挡情况，能够更好地评测模型对遮挡的鲁棒程度。本节接下来分别介绍了实验细节、与其他方法的对比实验以及消融实验。

5.4.1 实现细节

视频行人重识别的实现细节 本章采用Adam优化器 [95] 进行网络优化，共训练150个回合。初始学习率设为3×10^{-4}，每训练40个回合后下降10倍。输入图像的大小为256×128。训练使用的批量大小为32，其中包含8个行人类别，每个类别包含4个视频片段。在训练时，输入的视频片段的长度应该相同，并且每一帧应该以相同的概率被采样。因此，对于每个原始视频，以8帧为间隔随机采样4帧形成一个视频片段。本章使用水平翻转和随机擦除作为数据增广。超参数λ_1和λ_2和λ_3分别设置为0.1，0.5和0.05。

测试阶段，首先将测试序列等分为多个含有64帧的视频片段，然后利用训练完成的RFCnet提取每个视频片段的特征，最后取所有视频片段特征的均值作为输入视频的特征。对于每个查询视频，计算它与所有候选视频的特征相似性。根据计算出的特征相似性对所有候选视频进行排序，选出最相似的候选视频，达到目标检索的目的。

图像行人重识别的实现细节 在图像行人重识别任务上，RFCnet网络移除了TRFC模块。本章方法共训练60个回合，初始学习率设为3.5×10^{-4}，每训练20个回合后下降10倍。其他的训练设置与视频数据集上的一致。

5.4.2 消融实验

为了验证RFCnet网络各个模块的有效性，本节分别在Occluded-DukeMTMC-

表 5.1 在基于视频的遮挡行人重识别数据集上的消融实验。**Param**表示模型的参数量；**GFLOP**表示模型平均处理一帧的浮点运算次数；**TT**表示模型的训练时间；**IT**表示模型的推理时间

方法	Occluded-DukeMTMC-VideoReID					
	Param.	GFLOPs	TT	IT	mAP (%)	top-1 (%)
Baseline	23.5M	4.06	250m	25m	69.3	68.9
Foreground	23.5M	4.06	255m	25m	74.1	73.2
SRFC	24.0M	4.21	270m	26m	**82.9**	**82.6**
SRFC-A	-	-	-	-	80.6	79.1
SRFC-P	-	-	-	-	82.4	81.9
TRFC	23.8M	4.20	270m	27m	82.2	82.2
S-T-RFC	24.2M	4.34	280m	28m	**89.4**	**89.4**
T-S-RFC	-	-	-	-	88.9	89.0
S+T-RFC	-	-	-	-	86.5	86.1
RFC ($stage_2$)	24.2M	4.34	280m	28m	89.4	89.4
RFC ($stage_3$)	26.2M	4.34	281m	29m	**91.1**	**91.5**
RFCnet-wo-L_k	-	-	-	-	91.1	92.9
RFCnet-wo-L_f	-	-	-	-	90.8	91.1
SRFCnet	-	-	-	-	83.1	82.7
RFCnet	26.9M	4.62	290m	31m	**92.0**	**93.0**

VideoReID视频数据集和Occluded-DukeMTMC图像数据集上采取了一系列消融实验。本节引入一个常用的基准模型Baseline。基准模型使用ResNet50作为特征提取器，通过交叉熵损失和三元组损失联合训练。为了公平比较，Baseline使用与RFCnet相同的训练细节。表5.1和5.2总结了不同设置下的比较结果。

SRFC模块的有效性　如表5.1和5.2所示，SRFC模块在视频和图像重识别任务上带来一致的性能提升。为了进一步说明特征补全的有效性，本节引入一个对比方法Foreground，该方法利用前景图直接丢弃遮挡区域。SRFC的性能显著优于Foreground，在Occluded-DukeMTMC和Occluded-DukeMTMC-VideoReID数据集上分别提升了4.7%和8.8%mAP。实验结果验证了相比于丢弃遮挡区域的方法，修复遮挡区域的补全操作能够更有效地处理遮挡问题。本节进一步比较了不同的SRFC模块：SRFC-A和SRFC-P。这两个模块分别使用外观分配矩阵和位置分

表 5.2 在基于图像的遮挡行人重识别数据集上的消融实验

方法	Occluded-DukeMTMC					
	Param.	GFLOPs	TT	IT	mAP (%)	top-1 (%)
Baseline	23.5M	4.06	72m	110s	45.8	53.9
Foreground	23.5M	4.06	73m	110s	47.1	55.9
SRFC	23.9M	4.20	75m	112s	**51.8**	**60.9**
SRFC-A	-	-	-	-	50.6	58.6
SRFC-P	-	-	-	-	51.3	60.0
RFC ($stage_1$)	23.7M	4.21	76m	112s	49.9	58.9
RFC ($stage_2$)	23.9M	4.20	75m	112s	51.8	60.9
RFC ($stage_3$)	25.2M	4.20	74m	111s	**52.8**	**61.9**
RFC ($stage_4$)	30.0M	4.63	80m	115s	49.5	58.2
RFCnet-wo-L_k	-	-	-	-	52.9	62.0
RFCnet-wo-L_f	-	-	-	-	52.5	61.4
RFCnet	25.6M	4.34	78m	119s	**54.5**	**63.9**

配矩阵作为编码矩阵。如表5.1和5.2所示，SRFC-A在视频和图像的遮挡数据集上分别提升了11.3%和4.8%mAP，SRFC-P同样取得了优于Baseline的性能。实验结果表明了引入外观信息和位置先验对于遮挡补全的有效性。通过结合外观和位置信息，SRFC可以取得更优的性能，说明了外观信息和位置信息的互补性。

TRFC模块的有效性 本节进一步评测了TRFC的有效性。如表5.1所示，TRFC的性能显著优于Baseline，在mAP和top-1上分别提升了12.9%和13.3%。性能的显著提升归功于TRFC的长期时序建模能力。TRFC能够关注远距离帧的视觉线索，使补全的特征更加精确，同时具有与视频一致的语义信息。RFC模块整合了SRFC和TRFC，在mAP和top-1上进一步提升了7.2%，说明了SRFC与TRFC的互补性。

SRFC与TRFC的组合方式 本节比较了三种SRFC和TRFC的组合方式：顺序排列SRFC和TRFC（S-T-RFC），顺序排列TRFC和SRFC（T-S-RFC）以及并行使用两个模块（S+T-RFC）。表5.1总结了不同组合下的结果，可以观察到顺序组合方式能够取得高于并行组合的性能。这是由于顺序组合支持渐进式遮挡修复：其中一个模块给出一个粗糙的初步预测，另一个模块再对初步预测结果进行精细化。值得注意的是，所有组合方式都优于单独使用SRFC和TRFC的性能，说

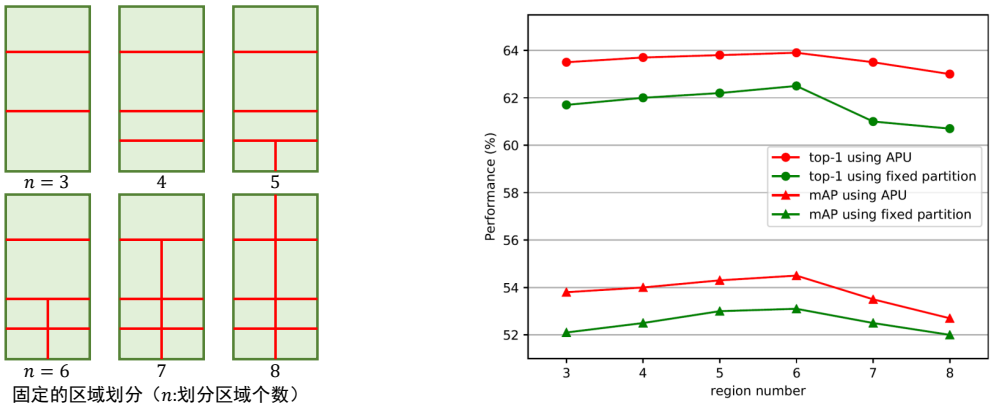


图 5.6 APU使用固定划分和自适应划分下在Occluded-DukeMTMC数据集的性能

明了同时使用空间和时间补全的重要性，并且最佳的组合方法能够进一步提高重识别性能。

RFC模块的放置位置　本节进一步测试了RFC的放置位置。如表5.2所示，RFC模块能够一致提升Baseline性能，并且放置在第二个和第三个阶段更为有效。一个可能的解释是：浅层特征的表现力较弱，不足以提供精确的语义线索；最后阶段的视觉概念过于抽象，难以传播有效的空间线索。本节进一步测试了放置多个RFC模块的性能。如表5.1和5.2所示，更多的RFC模块能够持续提升重识别性能。这是由于多个RFC模块形成了一个层次结构，其中后续RFC模块能够在之前模块修复的基础上预测被遮挡的特征，降低了遮挡修复的难度。

APU模块的影响　本节测试了APU的有效性，引入了一个APU的变体，该变体使用固定划分将特征图划分为多个固定的区域。如图5.6所示，APU一致优于固定划分的性能，说明了自适应划分策略的优越性。此外，APU在划分区域个数为6时取得了最优性能。这是由于太少的划分区域个数无法准确定位到遮挡部位，而太大的划分区域个数可能将非遮挡部件划分为多个区域，破坏了部件的语义信息。

关键点约束L_k的有效性　本节测试了关键点约束L_k的有效性，引入了一个比较模型RFCnet-wo-L_k，该模型移除了等式5.13中的关键点L_k损失。如表5.1和表5.2所示，不使用L_k会导致性能下降，说明了关键点约束的有效性。图5.7(a)可视化了RFCnet和RFCnet-wo-L_k生成的区域划分结果。从图中可以观察到通过引入关键点约束，RFCnet能够使每个划分区域集中到一个指定的身体部件。相反，RFCnet-wo-L_k的区域划分是杂乱无章的，难以精确定位到遮挡部件。

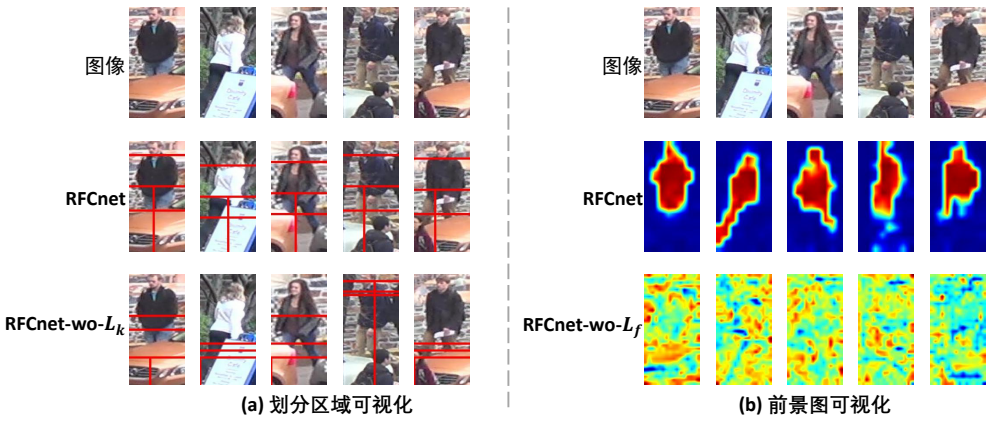


图 5.7 (a) 可视化不同模型生成的的划分区域结果（b）可视化不同模型生成的前景图

图 5.8 Occluded-DukeMTMT和Occluded-DukeMTMC-VideoReID数据集中同一个行人的查询图像和查询视频

前景图约束L_f的有效性 本节测试了前景图约束L_f的有效性，引入了一个比较模型RFCnet-wo-L_f。RFCnet-wo-L_f移除了等式5.13中的前景图约束L_f。如表5.1和5.2所示，RFCnet的性能一致优于RFCnet-wo-L_f，验证了前景图约束的有效性。图5.7(b)可视化了RFCnet和RFCnet-wo-L_f生成的前景图。可以观察到，RFCnet的前景图能够精确定位到行人的可视身体区域。相反，RFCnet-wo-L_f的前景图通常会集中到背景以及遮挡区域。因此RFCnet-wo-L_f很难消除遮挡物的干扰，导致性能下降。

计算开销比较 为了说明RFC模块的计算开销，本节报告了模型的参数量（Params），浮点运算次数（GFLOPs），训练时间（TT）以及推理时间（IT）。如表5.2所示，RFCnet引入了2.1M参数量，相比于Baseline只增加了8.9%。在计算开销上，RFCnet处理一张图像需要4.34GFLOPs，相比于Baseline只增加了6.8%。在时间开销上，RFCnet需要78分的训练时间和119秒的推理时间，比Baseline分别增加了8.3%和8.2%。综上，RFC模块只引入不到10%的计算和时间开销，能够应用到实时场景中。

表 5.3 本章方法与现有视频行人重识别方法在Occluded-DukeMTMC-VideoReID数据集性能对比

	方法	Occluded-DukeMTMC-VideoReID			
		mAP	top-1	top-5	top-10
集合建模	TriNet [21]	64.1	63.1	82.6	87.4
	STAN [46]	69.4	69.4	88.1	91.7
	QAN [44]	74.8	75.1	90.6	93.4
	RQEN [45]	74.9	73.5	90.5	94.1
时空建模	RCN [48]	62.4	60.9	83.3	88.1
	Snippet [61]	72.5	74.1	89.2	92.1
	IANet [68]（第二章）	81.2	81.9	93.9	95.6
	TCLNet [120]（第三章）	82.1	82.0	93.5	96.2
图像补全	VRSTC [121]（第四章）	82.7	84.0	94.4	97.1
特征补全	**RFCnet**（本章方法）	**92.0**	**93.0**	**98.6**	**99.1**
特征补全+图像补全	**RFCnet+VRSTC**	**92.1**	92.9	98.7	98.9
特征补全+完备特征学习	**RFCnet+TCLNet+IA**	**94.0**	**94.1**	**98.9**	**99.3**

5.4.3　视频和图像的性能分析

如表5.1和5.2所示，RFCnet在视频数据集（Occluded-DukeMTMC-VideoReID）上的性能显著高于同源的图像数据集（Occluded-DukeMTMC）上的性能。这一现象主要有两个原因。首先，视频序列的非遮挡帧能够部分解决遮挡问题。图5.8展示了 Occluded-DukeMTMC和OccludedDukeMTMC-VideoReID中同一个行人的查询图像和查询视频。在图像数据集上，由于行人的下半身被完全遮挡，模型只能根据行人的上半身识别其身份。在视频数据集上，行人的整体外观可以通过视频的非遮挡帧获取。因此，视频的非遮挡帧减轻了遮挡造成的信息损失。如表5.1和5.2所示，Baseline在视频数据集上的性能比在图像数据集上的性能提升了15%，验证了输入序列中更多的行人帧能够部分解决遮挡问题。其次，TRFC模块在视频行人重识别任务上带来了额外的性能提升。如图5.8所示，行人的下半身被完全遮挡。由于缺乏视觉线索，SRFC很难恢复遮挡区域的特征表示。相反，TRFC能够利用来自非遮挡帧的可视信息预测遮挡帧的特征，从而取得更优的性能。如表5.1和5.2所示，SRFC在视频和图像数据集上分别提升了13.8%和10%top-1，TRFC在视频数据集上进一步提升了10.3%top-1。实验结

表 5.4 本章方法与现有视频行人重识别方法在MARS和DukeMTMC-Video数据集性能对比

	方法	MARS		DukeMTMC-VideoReID	
		mAP (%)	top-1 (%)	mAP (%)	top-1 (%)
集合建模	STAN [46]	65.8	82.3	-	-
	EUG [76]	67.4	80.8	78.3	83.6
	RQEN [45]	71.7	77.8	-	-
	TAFD [47]	78.2	87.0	-	-
时空建模	M3D [52]	74.1	84.4	-	-
	Snippet+OF* [61]	76.1	86.3	-	-
	GLTP [94]	78.5	87.0	93.7	96.3
	CoSAM [97]	79.9	84.9	93.7	96.2
	IANet [68]	85.0	90.2	96.1	96.9
	TCLNet [120]	85.1	89.8	96.2	96.9
图像补全	VRSTC [121]（第四章）	82.3	88.5	93.8	95.0
特征补全	**RFCnet**（本章方法）	**86.3**	**90.7**	**97.0**	**97.6**

果验证了TRFC带来了视频行人重识别上额外的性能增益。

5.4.4 与当前先进方法的对比

首先在视频行人重识别任务上将RFCnet与当前先进的方法进行比较。表5.3总结了本章方法与已有方法在视频遮挡数据集Occluded-DukeMTMC-VideoReID上的对比结果。可以观察到：（1）在基于集合建模和时空建模的方法中，文献[21, 48, 61] 以相同的权重处理每一个原始帧，忽略了遮挡的不利影响。RFCnet取得了远超这些方法的性能，在mAP和top-1上分别提升了27.7%和29.6%。实验结果表明非遮挡重识别方法难以直接应用到遮挡场景中；（2）在基于集合建模的方法中，文献 [44–46] 使用丢弃遮挡帧的方式缓解遮挡的影响。RFCnet相比于这类方法提升了19.5%top-1和15.2%mAP。实验结果表明了特征补全相对于丢弃遮挡帧机制的有效性和优越性。表5.4总结了本章方法与已有方法在数据集MARS和DukeMTMC-VideoReID上的对比结果。RFCnet与众多方法相比都达到了最高水平，进一步表明了特征补全的有效性。

然后在图像行人重识别任务上将RFCnet与当前先进的方法进行比较。表5.5总

表 5.5 本章方法与现有图像行人重识别方法在Occluded-DukeMTMC数据集性能对比

方法	Occluded-DukeMTMC			
	mAP (%)	top-1 (%)	top-5 (%)	top-10 (%)
Part Aligned [122]	20.2	28.8	44.6	51.0
SFR [123]	32.0	42.3	60.3	67.3
Part Bilinear [90]	-	36.9	-	-
HACNN [124]	26.0	34.4	51.9	59.4
Random Erasing [107]	30.0	40.5	59.6	66.8
DSR [111]	30.4	40.8	58.2	65.2
Adver Occluded [110]	32.2	44.5	-	-
PCB [14]	33.7	42.6	57.1	62.9
PCB+RPP [14]	35.0	46.8	61.1	67.3
FD-GAN [125]	-	40.8	-	-
PGFA [22]	37.3	51.4	68.6	74.9
RFCnet （本章方法）	**54.5**	**63.9**	**77.6**	**82.1**

结了在图像遮挡数据集Occluded-DukeMTMC上的对比结果。从表5.5可以观察到：（1）本章方法的性能显著优于不考虑遮挡因素的行人重识别方法 [14, 122, 125]，在top-1上大约提升了20%。实验结果说明遮挡很大程度干扰了非遮挡行人重识别方法的特征提取过程。（2）遮挡行人重识别方法PGFA [22] 通过丢弃遮挡区域减弱遮挡物的干扰。相比于PGFA [22]，RFCnet的top-1提升了12.5%，mAP提升了17.2%。性能的大幅提升是由于特征补全有利于区分具有相似可视部件的不同身份行人。此外，PGFA只在最后一个网络层处理遮挡。RFCnet在不同层嵌入多个RFC模块，能够从底层逐步缓解遮挡的不利影响，进一步提高模型对遮挡的鲁棒性。（3）文献 [107, 110] 使用数据增强策略，通过生成更多的遮挡图像提高模型处理遮挡的能力。相比于数据增强的方法，RFCnet大约提升了2%mAP。值得注意的是，作为一种数据增强技术，上述方法 [107, 110] 能够与本章方法结合，进一步提高遮挡场景下重识别精度。

5.4.5　与第四章方法VRSTC的对比和结合

上一章的VRSTC [121] 使用一个昂贵的图像补全网络预测遮挡区域的像素，造成行人重识别框架过于复杂和耗时。相比之下，本章的RFCnet引入了极少的

计算开销和参数，同时进一步提升了重识别性能。如表5.3所示，在Occluded-DukeMTMC-VideoReID数据集上，RFCnet比VRSTC提升了13.4%mAP和15.3%top-1。如表5.4所示，在MARS和DukeMTMC-Video数据集上，RFCnet在mAP和top-1上大约提升了4%。实验结果表明了特征补全相比于图像补全的优越性，特别是在遮挡场景下特征补全的提升更为显著。本节进一步测试了结合特征补全和图像补全方法的性能。如表5.3所示，结合图像补全并不会带来性能提升，说明特征补全已经覆盖了图像补全学到的知识。

5.4.6 与第二、三章完备特征学习方法的结合

针对行人重识别的难点问题，本文分别提出了两类方法：特征完备学习（第二章和第三章），以及遮挡鲁棒学习（第四章和第五章）。本节测试了结合这两类方法的重识别性能。具体实现上，我们将TCLNet+IA作为主干网络，并在主干网络中嵌入遮挡鲁棒的特征补全模块RFC。如表5.3所示，在Occluded-DukeMTMC-VideoReID数据集上，RFCnet+TCLNet+IA比RFCnet提升了2.0%mAP。实验结果表明了特征完备和遮挡鲁棒学习能够相互整合，进一步提升重识别性能。因此，一个重识别系统可以同时使用这两类方法：一方面，TCLNet+IA能够为非遮挡的视频序列提取一个更完备的特征表示；另一方面，RFC能够避免遮挡序列的特征损坏，联合这两类方法可以获得更高的重识别性能。

5.5 本章小结

本章提出在特征层面补全遮挡区域内容。在特征补全中，遮挡修复和行人重识别任务相互耦合，因此行人重识别的反馈信号能够直接监督遮挡修复操作，充分提升重识别性能。同时特征补全模块只增加了不到10%的计算开销，能够嵌入到任意已有行人重识别方法中，提高已有模型对遮挡的鲁棒性。在多个数据集下，本章所提方法均取得了最高的性能，验证了本章方法的有效性。在未来的工作中，如何将本章方法应用到遮挡场景下的其他视觉任务，例如跟踪和检测，是一个值得研究的问题。

图像补全和特征补全都是针对遮挡场景所提出的方法。这两个方法具有相同的核心思想，即利用视频的时空结构性修复遮挡区域的内容。这两类方法具有不同的优缺点。（1）图像补全作为一个数据前处理方式，与重识别模型是互相解耦的。因此它的优势是能够与任意已有的行人重识别模型结合，无需再

训练重识别模型；它的缺点在于引入了一个昂贵的生成网络，时间开销比较大。（2）特征补全作为一个特征提取的模块，需要与重识别网络端到端训练。因此它的优势是能够使用重识别的反馈信号能够直接监督遮挡修复操作，产生更高的重识别精度，并且参数量和时间开销都比较小；它的缺点是改变了重识别模型的结构，需要重新训练一个重识别模型。基于上述的优缺点，可以根据不同的实际需求使用不同的补全方法。

第6章　高效时空建模框架

前几章分别从重识别模型的准确性和鲁棒性出发，本章并列探究了重识别模型的计算效率问题。本章设计了一个高效时空建模框架，在提升精度的同时显著提高了重识别的计算效率。

6.1　引言

目前，基于深度网络的视频行人重识别方法的性能得到不断提升。但是对于最新的模型（包括本文前几章方法），通过引入额外的模块和更多的参数来提高精度是很常见的。例如，最近的MGH [66] 添加了一个多尺度图卷积模块，增加了大约10%的计算量；本文前几章方法也增加了不同程度的计算开销。但是，例如机器人、自动驾驶汽车等移动设备无法部署大型且耗电的强大GPU，因此有限的计算资源阻碍了最新的重识别模型的实际应用。这启发我们开发在准确率、参数和计算效率取得良好平衡的视频行人重识别模型。

视频行人重识别模型的关键在于高效建模行人的空间和时序信息。首先，对于行人的空间外观信息，第三章指出现有大多数方法以相同的分辨率和网络结构处理每一帧图像，造成连续帧特征的高度相似和冗余。为了解决这个问题，第三章提出了时序互补建模思想：对连续帧挖掘互补的空间区域以减小相邻帧的特征冗余。但是第三章设计的TSE具有以下局限性：（1）TSE在计算上是不高效的。TSE需要多个顺序的学习器为连续帧挖掘互补的空间区域，因而TSE无法并行处理连续帧，增加了时间开销。（2）TSE引入了更多的参数量。TSE的多个学习器增加了6.4M的参数量，占基准模型的27%。综合上述分析，构建一个更高效轻量的空间建模模型是十分必要的。

其次，对于行人的时序动作信息，大多数方法只单独建模短期 [51, 80] 或长期 [48, 58] 的时序关系。为了增强时序建模能力，最近工作 [52] 尝试联合捕捉视频的短期和长期时序信息，并以相同的权重融合这两种时序关系。但是，短期和长期的时序关系对于不同性质的行人序列具有不同的重要度。如图6.1所示，对于一个部分遮挡的行人序列，长期的时序关系更有可能覆盖到非遮挡帧，更利于捕捉行人的完整运动模式；而对于一个快速移动行人的序列，短期的时序

图 6.1 短期和长期的时序关系对不同行人序列具有不同的重要性

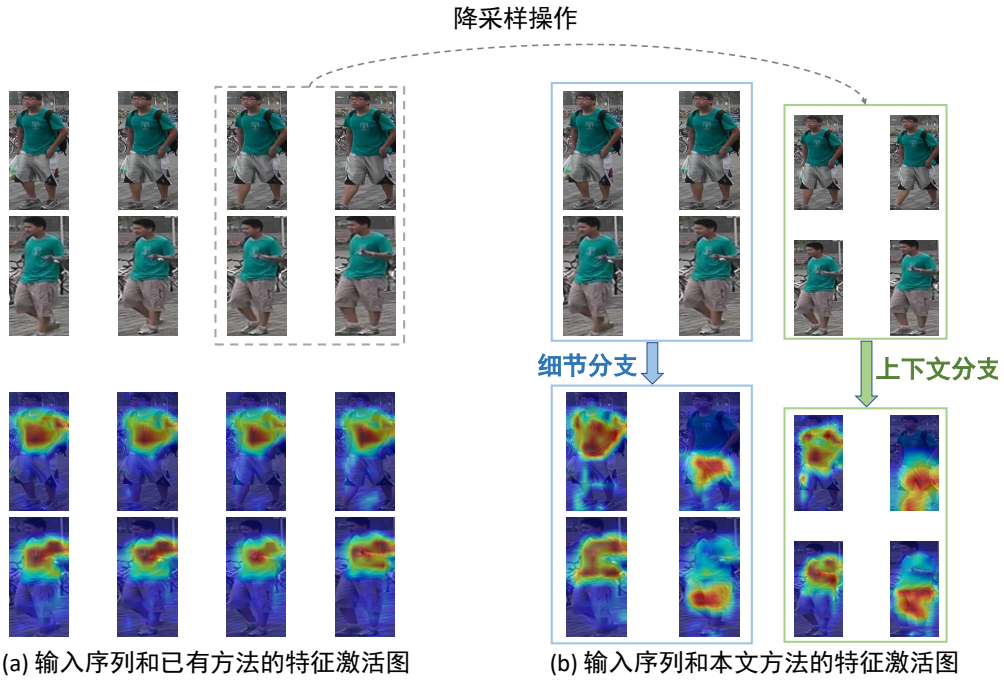


图 6.2 输入序列和模型特征激活图的示意图

关系更有利于建模行人的详细运动模式。因此，自适应捕捉短期和长期的时序关系是十分必要的。

为了实现上述目标，本章提出了一个高效时空特征表示框架。首先，本章提出了一个双分支互补网络（Bilateral Complementary Network，BiCnet），旨在高效提取连续帧的互补空间特征。本章观察到，降低图像的分辨率可以提高卷积特征图在原始图像的感受野，模型能够关注更远距离的空间上下文区域，同时使用低分辨率的输入帧大幅降低了计算开销。基于上述现象，BiCnet提出以不同的分辨率处理连续帧，达到高效互补建模的目的。如图6.2所示，BiCnet包含两个不同尺度的分支：细节分支和上下文分支。其中，细节分支处理原始分辨率的帧，用于保留细节视觉线索；上下文分支处理降采样后的低分辨率帧，用于增大网络的感受区域以捕捉到更远距离的身体信息。如图6.2所示，在更低

的输入分辨率下，第三帧的特征关注到一个更广泛的视觉线索（有/无背包背带的绿色T恤）。同时在每个分支上，BiCnet添加了多个并行的空间注意力模块。通过约束注意力模块之间的离散性，这些注意力模块能够关注到连续帧不同的空间区域。如图6.2所示，通过使用离散的注意力模块，相同分支上的连续帧可以关注不同位置的区域，从而覆盖目标行人的完整身体区域。不同于TSE使用顺序学习器挖掘互补的区域，BiCnet使用并行且轻量的注意力模块以及共享的学习器，具有更少的参数量和计算开销。

在时序建模方面，本章设计了一个时序核选择模块（Temporal Kernel Selection，TKS），旨在自适应建模短期和长期的时序关系。本章观察到，在时间维度上同时使用小尺度和大尺度的卷积核可以联合捕捉短期和长期的时序关系。因此，TKS包含多个并行且具有不同卷积核大小的时间卷积路径。同时，TKS根据多个路径的全局信息选择一个占主导地位的卷积尺度。通过上述选择机制，TKS可以根据输入视频的性质自适应改变时序建模的尺度，具有更强的时序表示能力。TKS模块具有很少的参数和计算量，因而很容易嵌入到BiCnet，构成“BiCnet-TKS”，联合学习行人视频的时空模式。

本章剩余章节安排如下：章节6.2详细介绍了基于双分支的高效时空建模框架；章节6.3给出了本章方法在公开数据集上的有效性分析；章节6.4将对本章内容进行小结。

6.2　基于双分支的高效时空建模框架

本章的目标是为视频行人重识别任务构建一个高效的时空建模框架。该框架由两个组件构成：用于高效空间建模的双分支互补网络BiCnet；用于高效时序建模的时序核选择模块TKS。TKS能够嵌入到BiCnet的任意阶段，构建高效时空特征表示框架BiCnet-TKS。接下来分别对BiCnet和TKS进行详细介绍。

6.2.1　高效空间建模的双分支互补网络

在视频行人重识别任务中，大多数现有方法对每一帧执行相同的操作，导致连续帧的特征高度冗余。如图6.2(a)所示，这些冗余的特征趋向于关注同一个局部区域，难以区分局部相似的不同行人。第三章提出了一种多帧互补建模方法，其核心思想是对连续帧挖掘互补的空间区域，减小相邻帧的特征冗余。但是，第三章方法使用多个有序学习器，增加了模型的参数量和时间开销。

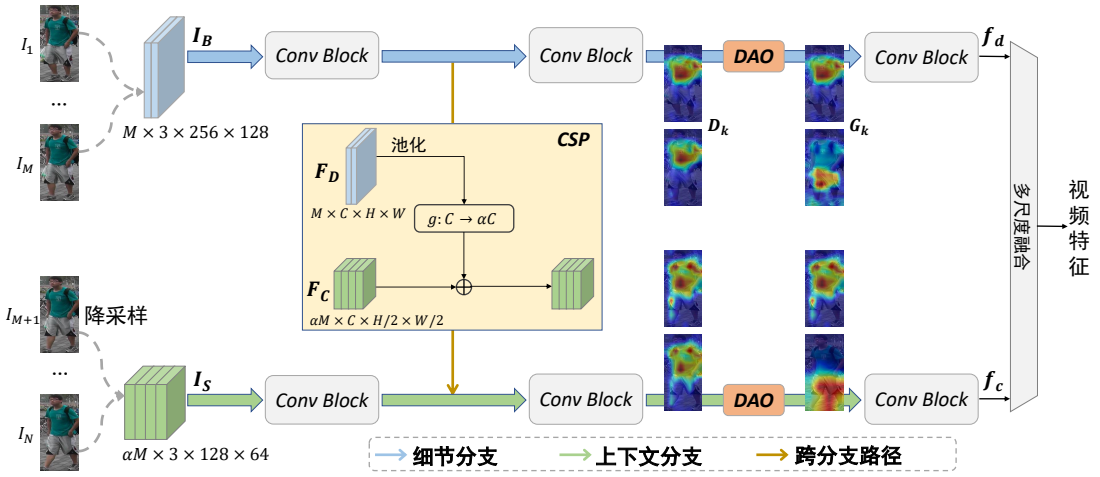


图 6.3 双分支互补网络（Bilateral Complementary Network，BiCnet）结构图

本章延续了多帧互补建模思想，设计了一个更高效的双分支互补网络BiCnet。BiCnet对连续帧关注不同尺度以及不同位置的空间区域，进一步提高了多帧互补建模的能力。

BiCnet的整体结构如图6.3所示。 BiCnet构建在一个双分支架构上，并在每个分支上添加一个离散注意力操作（Diverse Attentions Operation，DAO）。其中，双分支架构的目的是对不同的视频子片段建模互补尺度，离散注意力操作的目的是对连续帧挖掘不同位置的互补区域。通过在每个分支上添加DAO，BiCnet可以刻画目标行人的多尺度整体外观，提取一个更全面的空间表观特征。接下来，本节分别对BiCnet的双分支架构和离散注意力操作进行详细介绍。

6.2.1.1 双分支架构

如图6.3所示，BiCnet包含一个细节分支和一个上下文分支。给定一个输入的视频片段，细节分支以原始分辨率处理视频片段的前几帧，上下文分支以1/2的分辨率处理剩余帧。通过使用一个小尺度处理输入帧，上下文分支可以提供更大的感受区域，从而编码更远距离的空间上下文。因此，上下文分支与细节分支是功能互补的，其中细节分支集中捕捉行人的细粒度视觉线索，上下文分支集中捕捉粗粒度上下文信息。

具体实现上，给定一个包含N帧的视频片段，$I = \{I_n\}_{n=1}^{N}$，其中I_n表示视频片段的第n帧。首先将I划分为两个视频子片段，$I_B = \{I_n\}_{n=1}^{M}$和$I_S = \{I_n\}_{n=M+1}^{N}$。其中$I_B$的输入分辨率为原始分辨率；$I_S$的输入分辨率为原始分辨率的1/2；$M$决定了小帧与大帧数量的比例$\alpha$（$\alpha = \frac{N-M}{M}$）。然后将$I_B$和$I_S$分别送入到细节分支

和上下文分支中，提取对应的特征向量f_d和f_c。整个过程形式化如下：

$$\begin{aligned} f_d &= \frac{1+\alpha}{N}\sum_k \mathrm{CNN_D}(I_k), \quad I_k \in \{I_n\}_{n=1}^{\frac{N}{1+\alpha}} \\ f_c &= \frac{1+\alpha}{\alpha N}\sum_k \mathrm{CNN_C}(I_k), \quad I_k \in \{I_n\}_{n=\frac{N}{1+\alpha}+1}^{N}. \end{aligned} \tag{6.1}$$

最终取f_d和f_c的均值作为输入视频的特征表示。

6.2.1.2　跨分支路径

本节进一步引入一个跨分支路径（Cross-Scale Paths，CSP）将细节分支的中间层信息传播到上下文分支。通过使用跨分支路径，上下文分支能够感知到细节分支学到的特征，从而更专注于捕捉远距离的视觉线索。

CSP的结构如图6.3所示。给定细节分支和上下文分支相同卷积阶段的特征图$F_D \in \mathbb{R}^{M\times C\times H\times W}$和$F_C \in \mathbb{R}^{\alpha M\times C\times \frac{H}{2}\times \frac{W}{2}}$，其中$C$、$H$和$W$分别表示细节分支特征图的通道，高度和宽度。由于$F_D$和$F_C$具有不同的空间和时间维度，CSP首先将$F_D$转换为$\overline{F_D} \in \mathbb{R}^{\alpha M\times C\times \frac{H}{2}\times \frac{W}{2}}$，从而与$F_C$的大小相匹配：

$$\overline{F_D} = \mathcal{R}\left(W_c * \mathcal{P}(F_D)\right), \tag{6.2}$$

其中$\mathcal{P}$表示步长为2的空间池化操作以匹配F_C的空间维度；$*$表示卷积操作，$W_c \in \mathbb{R}^{1\times 1\times C\times \alpha C}$为卷积操作的核参数；$\mathcal{R}$为调整操作，将$M\times \alpha C\times \frac{H}{2}\times \frac{W}{2}$大小的卷积特征调整为$\alpha M\times C\times \frac{H}{2}\times \frac{W}{2}$大小，以匹配$F_C$的时间维度。最后，$\overline{F_D}$通过逐元素相加融合到$F_C$中。

6.2.1.3　离散注意力操作

如图6.3所示，虽然细节分支和上下文分支能够提供尺度互补的视觉线索（T恤和远距离的背带），但是每个分支上的连续帧会集中关注一个最具代表性的局部区域（上衣）。为此，本章设计了一个离散注意力操作DAO，旨在为连续帧挖掘不同位置的互补区域。通过在每个分支上添加一个DAO，BiCnet可以发现更丰富的空间区域，生成一个更完整的行人表观特征。

DAO的结构如图6.4所示。DAO由多个并行的空间注意力模块构成，并为每一帧使用一个指定的注意力模块。通过鼓励生成的注意力图之间的离散性，并行的注意力模块能够关注到不同位置的空间区域，从而为连续帧提取离散的判别特征。 DAO在细节分支和上下文分支上执行相同的操作。本节以细节分支

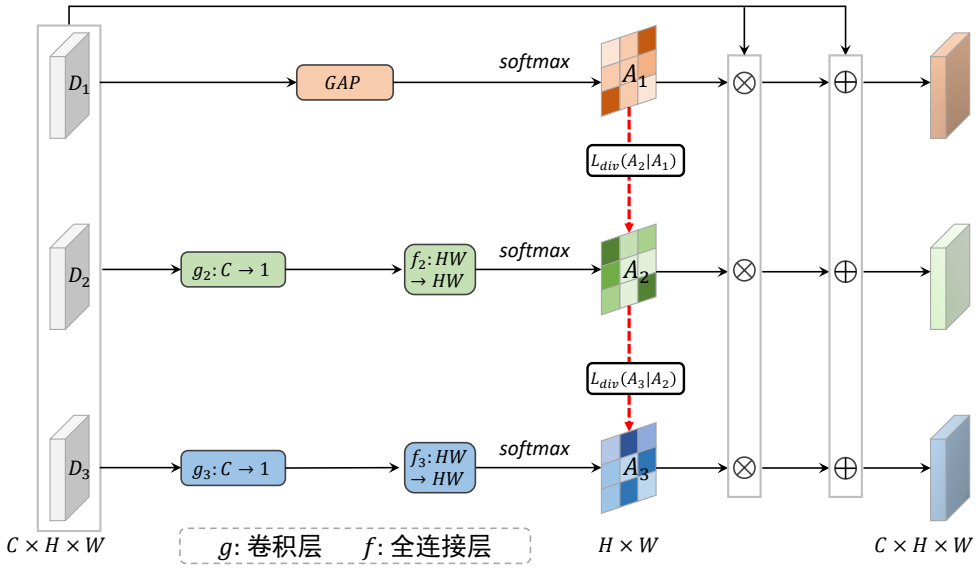


图 6.4 离散注意力操作（Diverse Attentions Operation，DAO）的结构图

为例说明DAO的工作原理。具体来说，给定细节分支上的特征图F_D，DAO为每帧$(F_D)_k \in \mathbb{R}^{C\times H\times W}$使用一个指定的注意力模块。为了简单起见，本节将$(F_D)_k \in \mathbb{R}^{C\times H\times W}$记做$D_k$。文献 [26] 指出，卷积特征图中每个特征单元的值与其判别性呈正相关。因此，DAO使用平均池化在通道维度上压缩D_1，以定位D_1激活的空间区域，产生一个空间注意力图$A_1 \in \mathbb{R}^{H\times W}$:

$$A_k = softmax\left(\frac{1}{C}\sum_{c=1}^{C}(D_k)_c\right), \quad k = 1. \tag{6.3}$$

然后，DAO使用多个并行的注意力模块去挖掘不同且未被激活的区域。具体来说，给定细节分支上第k帧的特征图 $D_k\,(k>1)$，第k个注意力模块首先使用一个卷积层压缩D_k的通道维度，并将卷积结果调整为HW大小；之后使用一个全连接层编码全局空间上下文信息；最后使用Softmax函数生成第k帧的空间注意力图$A_k \in \mathbb{R}^{H\times W}\,(k>1)$。

为了引导不同的注意力模块激活不同位置的空间区域，不同注意力模块生成的空间注意力图应该尽可能离散。为了实现这个目标，本节引入了一个离散正则化项衡量两个注意力图的离散度，定义如下：

$$L_{div}(A_k|A_l) = 1 - A_k^T A_l. \tag{6.4}$$

基于这个离散正则项，离散注意力损失的定义为：

$$L_a = \frac{-1}{M-1}\sum_{k=2}^{M}\left(\frac{1}{k-1}\sum_{l=1}^{k-1}L_{div}\left(A_k|A_l\right)\right). \tag{6.5}$$

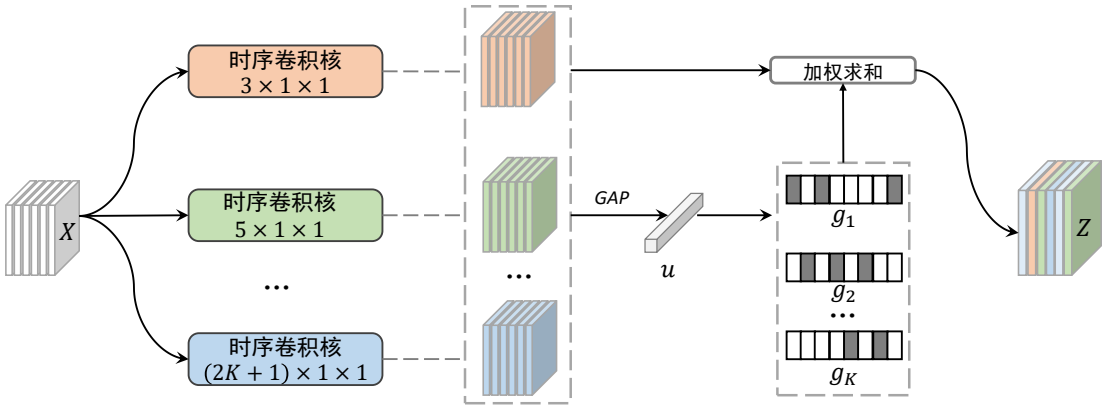


图 6.5 时序核选择模块（Temporal Kernel Selection，TKS）结构图

当任意两个注意力模块关注到相近位置的空间区域时，对应的注意力图会具有比较低的离散值，从而产生一个较高的损失值L_a。因此，使用L_a优化DAO可以驱使不同的注意力模块去关注不同位置的行人区域。

最后，DAO使用一个残余操作，将离散注意力信息编码到输入的特征图中。更新的特征图随后被送入对应分支后续的卷积层中，生成嵌入了互补视觉线索的特征向量。

6.2.2　时序核选择模块

受到视频分类任务 [49] 的启发，本方法将视频建模分解为空间建模和时间建模两个部分。为了建模视频的时序动作信息，本章设计了一个时序核选择模块TKS，联合建模短期和长期的时序关系。如图6.1所示，对于不同的行人序列，不同尺度的时序关系具有不同的重要程度。因此TKS提出以动态方式结合多尺度的时序关系，即根据输入的视频序列自适应对不同的时序尺度分配不同权重。

TKS的结构如图6.5所示。给定连续N帧的卷积特征图$F = \{F_n\}_{n=1}^{N}$，其中F_n为第n帧的特征图，TKS执行一系列操作：划分、选择和执行。接下来分别对每个操作予以介绍。

划分操作　文献 [51] 指出，由于行人重识别的数据来源于检测算法输出的行人裁剪图像，当检测算法存在误差时（检测框偏移、过大或者过小），相邻帧之间会出现空间不对齐，导致时间卷积对于视频行人重识别任务是无效的。针对这个问题，TKS利用划分策略对粗粒度的区域特征执行时间卷积，以缓解相邻帧的空间不对齐。具体实现上，给定$\{F_n\}_{n=1}^{N}$，TKS将每一帧特征图均匀划分

为$h \times w$个空间区域，并对每个划分的区域执行平均池化操作，构建一个区域级别的视频特征图$X \in \mathbb{R}^{N \times C \times h \times w}$。

选择操作 给定划分操作输出的X，TKS使用K个并行的时间卷积路径$\{\mathcal{F}^{(i)} : X \rightarrow Y^{(i)} \in \mathbb{R}^{N \times C \times h \times w}\}_{i=1}^{K}$，其中$\mathcal{F}^{(i)}$表示核大小为$2i+1$的时间卷积。为了提高计算效率，TKS将大小为$(2i+1) \times 1 \times 1$的卷积核替换为一个大小为$3 \times 1 \times 1$，膨胀因子为$i$的空洞卷积核。选择操作的核心思想是基于所有时间路径的全局信息决定分配给每个路径的权重。具体实现上，首先使用逐元素相加操作融合所有路径的输出信息，然后执行全局平均池化操作生成一个全局特征：

$$u = GAP_{N,h,w}(\sum_{i=1}^{n} Y^{(i)}), \tag{6.6}$$

其中$GAP_{N,h,w}$表示沿着时间和空间维度的全局平均池化操作。之后，根据全局编码生成每个时间路径的通道选择权重：

$$g_i = \frac{\exp(W_i u)}{\sum_{j=1}^{K} \exp(W_j u)} \quad i \in \{1, \ldots, K\}, \tag{6.7}$$

其中$W_i \in \mathbb{R}^{C \times C}$是生成$g_i$的转换参数。最后，将选择权重作用到对应的时间卷积核上，加权求和获得特征图$Z \in \mathbb{R}^{N \times C \times h \times w}$：

$$Z = \sum_{i=1}^{K} \mathcal{R}(g_i) \odot Y^{(i)}, \tag{6.8}$$

其中$\mathcal{R}$为调整操作，将$g_i \in \mathbb{R}^{C \times 1}$调整为$\mathbb{R}^{1 \times C \times 1 \times 1}$，以与$Y^{(i)}$的大小兼容。相比于使用单个标量权重提供粗粒度融合，TKS使用通道向量权重实现细粒度融合。融合的权重根据输入视频的性质动态决定，因此比较适用于不同序列具有不同主导时间尺度的视频行人重识别任务。

执行操作 根据选择操作输出的Z，执行操作使用残差机制调节输入特征图。最终的特征图$E = \{E_n\}_{n=1}^{N}$计算为：

$$E_n = \mathcal{U}(Z_n) + F_n, \tag{6.9}$$

其中$\mathcal{U}$表示基于最近邻的上采样器，对Z_n进行上采样以匹配F_n的空间分辨率。TKS模块维持了输入特征图的大小，可以嵌入到BiCnet的任意深度中，高效提取行人序列的时空特征。

6.2.3　网络结构和目标函数

网络结构　本章的高效时空特征建模框架是一个通用的框架，可以兼容任意的主干网络。BiCnet-TKS的分支构建在ResNet50上，每个分支包含四个连续的卷积阶段。CSP添加在每个卷积阶段之后。由于高层特征图具有更明确的语义信息，DAO添加到第三个卷积阶段之后。多分支架构带来的一个直接问题是使用多个分支会引入几倍的参数量，增加了模型过拟合的风险。针对这个问题，BiCnet提出共享两个分支的结构和参数，有效减小了参数的数量，并且具有与单分支行人重识别模型相同量级的参数。

计算开销　为了说明BiCnet-TKS的计算开销，本节考察一个常用的视频行人重识别基准模型Baseline。Baseline使用ResNet50作为特征提取器，以原始分辨率为每一个视频帧提取特征。假定ResNet50处理一个原始帧的浮点运算次数（GFLOPS）为p，给定一个包含N帧的视频序列，Baseline需要的计算量为Np。BiCnet-TKS以$1:\alpha$的比率将输入视频划分为原始分辨率帧和降采样分辨率帧。由于与特征提取相比，CSP，DAO和TKS的计算量可以忽略不计，因此BiCnet-TKS需要的计算量大约为$\frac{N}{1+\alpha}p+(\frac{\alpha N}{1+\alpha})\frac{p}{4}$。相比于Baseline，BiCnet-TKS减小了$\frac{3}{4}-\frac{3}{4\alpha+4}$的计算量。

从上述分析可以发现，BiCnet-TKS的计算量随着α的增大而降低。但是当α过大时，低分辨率帧会主导模型优化，造成性能严重下降。实验表明当α取值为3时，模型能够在计算成本和精度之间取得一个良好平衡。此时，BiCnet-TKS的计算量仅仅是Baseline的44%，可以更有效地为视频序列提取时空特征。

损失函数　训练BiCnet-TKS的目标函数包括：行人身份监督的交叉熵损失L_{ce}，困难样本三元组损失L_{tri}，以及离散注意力损失L_a（等式6.5）。其中交叉熵损失和三元组损失的细节见章节2.2.3.1。整体目标函数定义为：

$$L_{all}=(L_{ce}+L_{tri})+\lambda_1 L_a. \tag{6.10}$$

λ_1是用来平衡各个损失的超参数。

6.3　实验分析

本节分别在MARS [43]、 DukeMTMC-VideoReID [76] 和LS-VID [94]三个视频数据集上对本章提出的BiCnet-TKS进行评测。本节接下来分别介绍了实验细

节、与其他方法的对比实验、消融实验以及可视化分析。

6.3.1 实现细节

本章采用Adam优化器进行网络优化，共训练150个回合。初始化学习率设为3.5×10^{-4}，每训练40个回合后下降10倍。训练使用的小批量大小为64，其中包含16个行人类别，每个类别包含4个视频片段。在训练时，输入的视频片段的长度应该相同，并且每一帧应该以相同的概率被采样。因此，对于每个原始视频，以4帧为间隔随机采样8帧形成一个视频片段。本章采用水平翻转和随机擦除作为数据增广。在BiCnet中，原始分辨率为256×128，降采样后的分辨率为128×64，降采样帧与原始分辨率帧的比率α设为3。在TKS中，时间卷积核的个数K设为2。离散注意力损失的权重设为0.01。

测试阶段，首先将每个测试视频等分为多个含有8帧的视频片段；然后利用训练完成的BiCnet-TKS提取每个视频片段的特征；最后将所有视频片段的特征取均值得到测试视频的特征。对于每个要查询的视频，计算它与所有候选视频特征的余弦相似性。根据计算出的特征相似性对所有的候选特征进行排序，选出最相似的候选视频，达到目标检索的目的。

6.3.2 与当前先进方法的对比

本节将BiCnet-TKS与已有视频行人重识别方法进行比较。对比结果见表6.1。在MARS、DukeMTMC-VideoReID和LS-VID数据集上，BiCnet-TKS与众多对比模型均达到了较高水平。与基于集合建模和时空建模的方法相比，BiCnet-TKS具有可比的性能和更少的计算开销。基于集合建模和时空建模的方法以相同的分辨率和网络处理视频序列的每一帧，忽略了相邻帧的特征冗余问题。BiCnet-TKS以不同的分辨率处理不同帧，并使用DAO关注连续帧不同位置的空间区域，形成一个更完整的多尺度行人特征。相比于已有先进的时空建模方法MGH[66]和第二章方法IANet[68]，BiCnet-TKS在MARS数据集上分别提升了0.2%和1.0%mAP。相比于IANet，BiCnet-TKS在LS-VID数据集上分别提升了5.9%mAP和4.5%top-1。这些实验结果验证了BiCnet-TKS时空建模的优越性。在DukeMTMC-VideoReID数据集上，BiCnet-TKS的性能略低于IANet。这是由于BiCnet-TKS只保留了1/4帧的原始分辨率，无法充分捕捉短视频的表观和动作信息，不利于短视频的特征建模。由于DukeMTMC-VideoReID数据集存在大量的短视频序列，BiCnet-TKS在

表 6.1 本章方法与现有视频行人重识别方法的性能对比

方法		MARS		DukeMTMC-Video		LS-VID	
		mAP	top-1	mAP (%)	top-1 (%)	mAP	top-1
集合建模	STAN [46]	65.8	82.3	-	-	-	-
	EUG [76]	67.4	80.8	78.3	83.6	-	-
	TAFD [47]	78.2	87.0	-	-	-	-
时空建模	M3D [52]	74.1	84.4	-	-	40.1	57.7
	Snippet+OF [61]	76.1	86.3	-	-	-	-
	V3D [49]	77.0	84.3	-	-	-	-
	GLTP [94]	78.5	87.0	93.7	96.3	44.3	63.1
	CoSAM [97]	79.9	84.9	93.7	96.2	-	-
	IANet [68]	85.0	**90.2**	96.1	**96.9**	69.2	80.1
	AP3D [51]	85.1	90.1	95.6	96.3	73.2	84.5
	MGH [66]	85.8	90.0	-	-	-	-
	TCLNet [120]	85.1	89.8	**96.2**	**96.9**	70.3	81.5
	BiCnet-TKS	**86.0**	**90.2**	96.1	96.3	**75.1**	**84.6**

该数据集上表现不佳。

与第三章方法TCLNet的比较. 相比于TCLNet，BiCnet-TKS在MARS和LS-VID数据集上具有更好的性能。特别是在LS-VID数据集上，BiCnet-TKS的mAP提升了4.8%，top-1提升了3.1%。性能的大幅提升验证了BiCnet-TKS具有更强的多帧互补建模能力。这是因为TCLNet只考虑单尺度的特征提取；而BiCnet-TKS使用两个不同分辨率的分支，为连续帧提取尺度互补的特征，因此得到一个更鲁棒的多尺度特征表示。此外，BiCnet-TKS的TKS模块有效捕捉了行人序列的短期和长期动作信息，能够帮助区分表观相似的行人，进一步提升重识别精度。

与已有方法计算开销的比较. 本节进一步比较了BiCnet-TKS与已有方法的计算开销。表6.1中的已有方法都使用ResNet50作为主干网络，并在主干网络上添加特别设计的集合建模或者时空建模模块。因此这些方法的计算量都高于ResNet50，需要大于4.1GFLOPS的计算量处理一个视频帧。TCLNet使用顺序学习器挖掘相邻帧的互补区域，因而无法并行处理相邻帧，增加了视频处理的时间开销。BiCnet-TKS通过降低输入帧的分辨率减小了时间和计算开销。BiCnet-TKS平均需要2.0GFLOPS处理一个视频帧，只占大多数现有方法50%左

表 6.2 BiCnet-TKS的消融实验。GFLOPs表示模型平均处理一帧的浮点运算次数。Param.表示模型的参数量

方法	MARS			
	GFLOPs.	Param.	mAP (%)	top-1 (%)
Baseline-S (128×64)	1.02	23.5M	80.7	87.4
Baseline-B (256×128)	4.08	23.5M	85.2	89.1
Two-branch (TB)	1.81	23.5M	84.3	89.6
TB+CSP	1.89	27.6M	**85.0**	89.6
TB+CSP+AO (wo L_1)	1.89	27.6M	85.2	89.3
TB+CSP+DAO (BiCnet)	1.89	27.6M	**85.6**	**89.8**
BiCnet-TK (fix-fusion)	1.91	29.1M	85.5	89.6
BiCnet-TKS	1.99	29.2M	**86.0**	**90.2**

表 6.3 分支个数对BiCnet性能的影响

分支个数	高度H			MARS			
	256	128	64	GFLOPs.	Param.	mAP (%)	top-1 (%)
1	✓			4.08	23.5M	**85.2**	89.1
		✓		1.02	23.5M	80.7	87.4
			✓	0.25	23.5M	64.1	77.4
2	✓	✓		2.55	23.5M	84.8	**89.4**
3	✓	✓	✓	1.76	23.5M	79.1	86.1

右的计算量。

6.3.3 BiCnet的消融实验

本节通过在MARS数据集上采取一系列消融实验验证了BiCnet各个模块的有效性。本节引入了一个常用的基准模型。基准模型使用ResNet50独立提取每一帧的特征，然后通过时序平均池化融合多帧特征得到视频特征。基准模型以相同的分辨率处理所有的输入帧，并且使用交叉熵和三元组损失联合优化。为了公平比较，基准模型使用与本章方法相同的训练细节。本章考虑了两个基准模型，Baseline-B和Baseline-S。其中，Baseline-B使用256×128的原始分辨率作为输入分辨率，Baseline-S使用128×64的降采样分辨率作为输入分辨率。消融实验结果如表6.2所示。

表 6.4 小帧与大帧划分比率α对基础双分支架构TB性能的影响

划分比率α	MARS			
	GFLOPs.	Param.	mAP (%)	top-1 (%)
0 (Baseline-B)	4.08	23.5M	**85.2**	89.1
1	2.57	23.5M	84.8	89.4
2	2.07	23.5M	84.4	**89.7**
3	1.81	23.5M	84.3	89.6
4	1.67	23.5M	83.8	89.5
+∞ (Baseline-S)	1.02	23.5M	80.7	87.4

分支个数的影响　本节测试了BiCnet分支个数的影响。通过将视频帧划分为多组，并且为每组使用指定的分辨率，可以将双分支架构拓展到多分支架构。为了公平比较，本节使用均匀划分策略，即将视频序列的所有帧等分为多组。实验结果如表6.3所示，可以观察到：（1）使用合适的输入低分辨率可以提供较高的重识别精度。相比于原始分辨率，采用128×64的输入分辨率时，基准模型的top-1只下降了1.7%。同时采用低分辨率输入大幅降低了计算开销：128×64分辨率减少了75%的计算量。（2）太小的输入分辨率会导致严重的性能下降。当输入分辨率为64×32时，基准模型的mAP降低了21.1%。这是由于当输入分辨率太小时，模型丢失了大量的行人细节线索。（3）三分支架构的性能低于双分支架构。这是由于输入分辨率为64×32的分支会干扰分支间共享参数的优化。基于上述分析，本章使用一个双分支架构，能够以更少的计算量达到与Baseline-B相当的性能。

不同分辨率帧划分比率的影响　本节调查了划分比率α（等式6.1）对双分支架构（TB）的影响，其中α表示视频序列中128×64分辨率帧与256×128分辨率帧的数量比率。实验结果如表6.4所示。可以观察到，随着α的增加，TB大幅度降低了平均处理一帧所需的计算量。但是TB的性能会随着α的增加而降低。这主要是由于TB的两个分支之间缺乏交互，因此一个分支很难学到去集中捕捉另一个分支忽略的线索。此外，由于低分辨率帧特征的判别力较低，直接使用低分辨率帧不可避免会削弱最终的视频特征。相比于$\alpha = 2$，$\alpha = 3$只带来少量的性能降低。考虑到计算复杂度，本章方法将α设置为3。

跨分支路径的有效性　本节测试了跨分支路径CSP对模型性能的影响。如

表 6.5 结合不同时间卷积核对TKS性能的影响

卷积核个数K	卷积核大小			MARS			
	K3	K5	K7	GFLOPs.	Param.	mAP (%)	top-1 (%)
1	✓			1.94	28.3M	85.1	89.9
		✓		1.94	28.3M	85.3	90.1
			✓	1.94	28.3M	85.5	89.8
2	✓	✓		1.99	29.2M	**86.1**	**90.3**
	✓		✓	1.99	29.2M	85.7	90.0
		✓	✓	1.99	29.2M	85.6	90.1
3	✓	✓	✓	2.04	30.0M	85.8	90.2

表6.2所示，加入CSP能够带来0.7%mAP的提升，并且只增加了4%的计算量。性能的提升归功于细节分支到上下文分支的信息传播增强了上下文分支的表示能力。此外，两个分支可以协同工作去挖掘互补的视觉线索：细节分支提取局部身体部件的细节特征；上下文分支更侧重远距离的上下文特征。因此整合两个分支可以得到一个更具表现力的视频特征。

离散注意力操作的有效性 本节分别测试了离散注意力操作中注意力模块和离散约束的影响。本节引入一个对比模型TB+CSP+AO，该模型使用多个并行的注意力模块，但是缺乏离散约束的指导。如表6.2所示，相比于TB+CSP，TB+CSP+AO只取得了0.2%mAP的增益。实验结果表明TB+CSP+AO不同的注意力模块捕捉到的视觉线索几乎相同。TB+CSP+DAO（BiCnet）提升了0.6%mAP，验证了离散约束的有效性。

6.3.4 TKS的有效性分析

时序核选择机制的有效性 本节测试了TKS的有效性。如表6.2所示，TKS提升了0.4%mAP和top-1，并且只增加了6%的计算量。性能的提升表明TKS模块和BiCnet网络是功能互补的。其中，BiCnet集中提取行人的表观特征，TKS重点提取行人的动作特征。为了验证TKS自适应选择机制的有效性，本节引入了一个时序核模块（Temporal Kernel，TK）。TK将等式6.8替换为$Z = \frac{1}{K}\sum_{i=1}^{K} Y^{(i)}$，即对多尺度时序卷积的结果取均值。如表6.2所示，TK模块并不会带来重识别性能的提升，表明TKS的提升归功于多尺度时序卷积之间的自适应选择机制。

时间卷积核个数的影响 本节测试了结合不同时间卷积核的影响。本节考

表 6.6 TKS放置位置对BiCnet-TKS性能的影响

TKS放置位置	MARS			
	GFLOPs.	Param.	mAP (%)	top-1 (%)
stage1	1.99	28.0M	85.3	90.1
stage2	1.99	29.2M	**86.0**	90.2
stage3	1.99	34.1M	85.7	**90.4**
stage4	2.29	53.5M	85.4	90.0
stage2+stage3	2.09	35.7M	85.8	90.3

虑三种时间卷积核，K3，K5和K7。其中K3表示核大小为$3\times1\times1$的卷积；K5表示核大小为$3\times1\times1$，膨胀因子为2的空洞卷积；K7表示核大小为$3\times1\times1$，膨胀因子为3的空洞卷积。实验结果如表6.5所示。可以观察到：（1）使用两个不同尺度的时间卷积核能够一致提升模型性能。表6.5的第二部分（$K=2$）的性能普遍高于第一部分（$K=1$），表明了同时建模短期和长期时序的有效性。（2）使用更多的时序卷积核（$K=3$）不会带来性能的进一步提升，说明两个时序卷积核足够捕捉到行人视频的时序信息。

TKS模块的放置位置　本节进一步测试了TKS模块放置位置的影响。如表6.6所示，将TKS放置在stage2和stage3能够带来接近的精度提升，但是将TKS放置在stage1和stage4会导致性能下降。一个可能的解释是stage1的卷积特征图表现力较弱，不足以提供精确的语义线索，导致TKS无法准确建模行人部件的动作信息；同时由于BiCnet在stage3关注连续帧的不同区域，使得stage4的帧与帧之间的特征缺乏连贯的时序关系，造成TKS无法在stage4提取有效的动作特征。从表6.6可以观察到添加更多的TKS模块并不会进一步提高重识别性能，因此本章只在BiCnet的stage3嵌入一个TKS模块用于视频行人重识别。

6.3.5　可视化分析

特征图可视化　本节比较了Baseline和BiCnet特征图的可视化。如图6.6所示，Baseline的特征只集中到一个局部区域（蓝色上衣），很难区分图中的不同行人。相反，BiCnet能够对连续帧挖掘互补的视觉线索。首先，BiCnet的上下文分支通过降采样输入帧能够关注更广泛的视觉线索。如图6.6(a)所示，输入片段第三帧的特征能够关注到一个更大的空间区域，即带有特定标志的绿色T恤；其

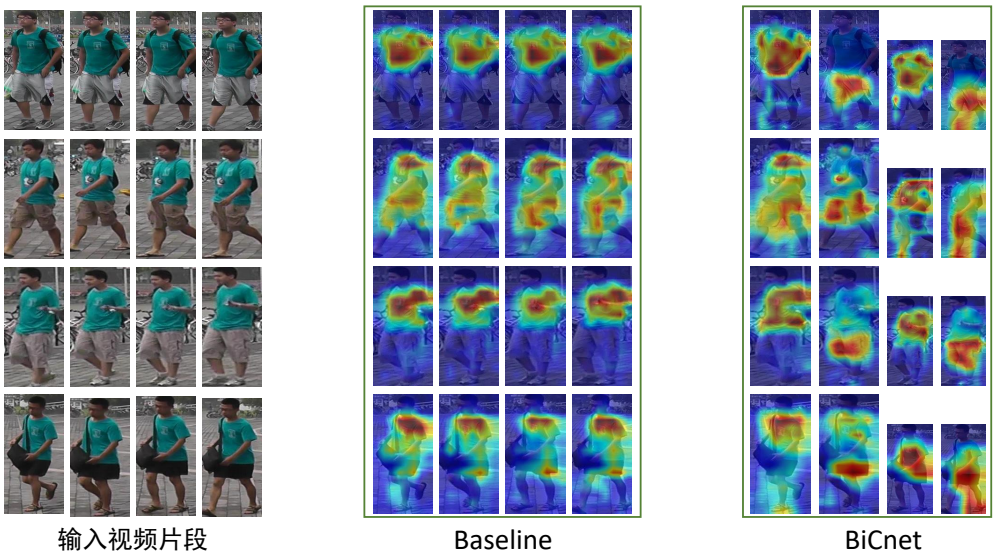


图 6.6 Baseline和BiCnet特征图的可视化

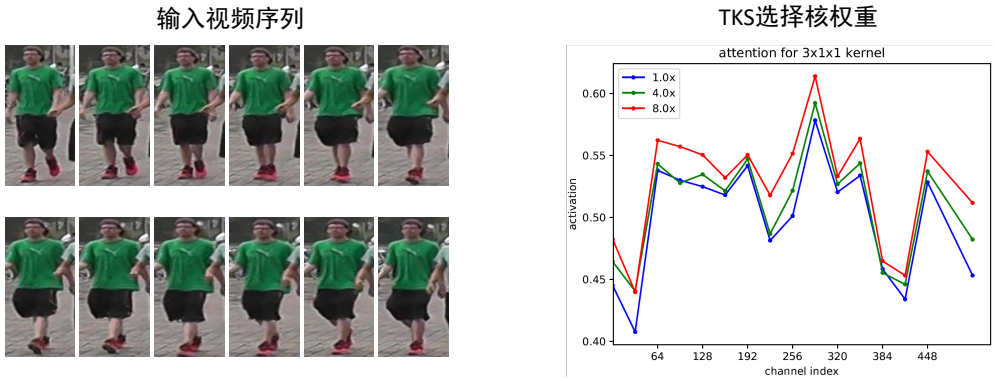


图 6.7 TKS选择权重的可视化

次，DAO能够对连续帧挖掘不同位置的空间区域。图6.6中第一帧和第三帧的特征均与绿色T恤相关，DAO的注意力模块在第二帧和第四帧中激活了行人的下半身区域，更容易区分这些局部相似的不同行人。

TKS的选择权重可视化 为了理解TKS自适应选择机制的工作原理，本节以不同时间步长τ采样一个原始视频，然后将采样后的视频序列输入到TKS中，可视化TKS生成的时序核选择权重。较高的τ会获得一个移动速度更快的行人序列。图6.7可视化了小尺度时序核（$3\times1\times1$）所有通道上的选择权重。可以观察到在大多数通道上，当目标行人移动更快时，小尺度时序核的选择权重会随之增加。可视化结果表明TKS自适应减小了时序关系的建模尺度，以捕捉详细的行人运动模式，验证了本章TKS的设计动机。

6.4　本章小结

针对视频重识别的效率问题，本章提出了高效时空建模框架。在空间维度上，本章设计了双分支互补网络，该网络通过以不同的分辨率处理不同的连续帧，高效提取到一个完整的多尺度表观特征；在时间维度上，本章设计了时序核选择模块，该模块通过自适应捕捉短期和长期的时序关系，提取到一个完整的多尺度动作特征。在大规模的视频行人重识别数据集的评测结果说明了双分支互补方式在效率上更具优越性，同时也说明了本章提出的时序核选择方法充分利用了视频的时序信息，可以显著提升重识别的性能。

未来的工作将考虑将本章方法应用到更多的视频任务中，比如视频人脸识别和视频行为分类等，以检验本章方法的通用性。

第7章　总结与展望

7.1　本文工作总结

本文针对视频行人重识别任务，以时空信息高效建模完备和鲁棒的行人特征的思路开展了以下五项工作：

时空交互的单帧判别特征学习方法　针对卷积操作受限于局部建模这一问题，提出了一个基于时空交互建模的特征学习方法。该方法通过时空交互操作生成一个时空关系图，有效编码行人部件之间的空间和时间关联。基于时空关系图在每帧特征中引入行人全局信息，提高了单帧特征的判别性，进而增强多帧特征融合后的视频特征。实验验证了全局建模思想对于行人重识别任务的有效性，而时空交互是实现全局建模的一个有效手段。然而，由于该工作以相同的操作处理每一帧图像，造成连续帧特征高度相似甚至冗余。这些高度相似的特征趋向于关注同一个显著但是局部的区域，导致最终的视频特征没有覆盖一个完整的行人身体，缺乏完整性。

时序互补的多帧完整特征学习方法　该工作独创性的观察到已有视频行人重识别方法存在连续帧特征冗余问题，导致难以处理行人重识别的类间相似问题。基于这个观察，提出了一个基于时序互补建模的特征学习方法。时序互补的核心思想是对连续帧提取互补的特征，使整合的视频特征能够覆盖一个更完整的行人身体区域。在时序互补建模的思想下，该工作设计了一个显著性擦除操作，通过显式擦除之前帧已经关注的区域，引导当前帧关注互补的区域。实验验证了，时序互补建模思想能够提高模型区分相似行人的能力，而显著性擦除操作是实现时序互补的一个有效手段。

面向遮挡场景的图像时空补全学习方法　针对遮挡造成行人信息损失这一问题，提出了一个利用时空信息进行遮挡补全的方法。该工作根据行人序列的空间结构和时间关联设计了一个时空补全网络，将重构学习、对抗学习和身份预测约束巧妙地联合到一起，进而形成了一个有效的图像层面的遮挡区域修复模型。实验证明了，在以部分遮挡为主导的条件下，时空补全思想相比于常用的丢弃遮挡帧策略更为有效。但是，图像补全的主要目标在于生成视觉真实的图像。因此，图像补全更加关注生成图像的视觉质量而非判别性，以至于补全

操作没有充分提升重识别性能。

面向遮挡场景的特征时空补全学习方法 在上一个工作的基础上，针对图像补全与重识别的目标不一致问题，提出了一个特征时空补全方法。在特征补全的方式下，遮挡修复任务和行人重识别任务相互耦合。此时，行人重识别模型的反馈信号直接指导遮挡修复模块的优化，使得补全操作能够充分提升重识别性能。同时，特征补全方法只增加了很少的计算开销，具有更好的实时性。实验证明了，相对于图像补全，特征补全方式能够以更少的计算成本和参数量，更优地提高行人重识别模型对遮挡的鲁棒性。

高效时空建模框架 针对视频行人重识别的效率较低这一问题，提出了一个基于双分支的高效时空建模框架。在空间特征学习上，通过以不同的分辨率分支处理不同的连续帧，提取一个多尺度表观特征，同时降低了计算开销。在时序特征学习上，通过自适应捕捉视频的短期和长期时序关系，学习一个多尺度动作特征。实验证明了，由于连续帧的特征冗余，以相同的时间间隔降低视频帧的分辨率的双分支框架能够保持模型的精度，同时降低了模型的计算量。因此，双分支结构是提高行人重识别效率的一个有效框架。

上述五个工作围绕视频行人重识别的时空建模，沿着特征更完备，遮挡更鲁棒，计算更高效的思路展开研究，并设计了相应的深度模型学习方法。这些方法从准确性、鲁棒性、高效性共同构成了实现视频行人重识别的综合性解决方案。所设计的方法在多个公开的行人重识别数据集上显示了良好的分析性能，并具有一定的实用价值。

7.2 关于方法局限性的讨论

尽管本文围绕基于视频的行人重识别进行了研究并取得了一定的成果，但是研究过程中仍然有一些问题没有完全解决。在本节中将对本文工作的局限性进行总结，并结合个人理解梳理出一些可能的解决方案。

面向特征学习的时序互补建模的基本思想是对连续的视频帧提取互补的特征。第三章的时序互补学习方法在后续帧中显式擦除判别性区域，可能会损害视频特征的表示能力。一个更好的策略是将硬擦除操作替换为软抑制操作，通过对之前帧关注的区域赋予一个低的权重引导模型挖掘更多的身体区域。另外，相对于时序互补建模，一些文献 [50, 97] 提出时序一致性建模，通过约束视频帧

之间的特征一致性减少低质量帧的干扰。时序互补和时序一致性思想的结合也是后续一个重要的研究方向。

面向遮挡问题的特征补全方法依赖行人姿态估计模型定位行人的各个身体部件区域。实际场景中存在严重行人遮挡的条件下，行人姿态估计模型可能失效，此时所提方法无法精确划分身体部件区域，降低了遮挡补全的有效性。一个可能的解决方案是不再依赖行人姿态估计模型，而是采用一种迭代的注意力机制，使得模型能够自动划分输入视频帧的遮挡区域与非遮挡区域。另外，严重遮挡的条件下，行人的非遮挡区域可能不足以提供足够的信息用于识别行人身份，此时可以结合人脸等不易发生遮挡的细粒度信息综合判断输入视频的行人身份。

面向高效建模的双分支网络采用固定的分帧策略。具体来说，给定任意一个包含N的视频序列，始终对前M帧保留原始分辨率，后续的$N-M$帧使用降采样分辨率。这种固定分帧方式忽视了不同帧的质量差异，以及不同序列的模式差异。如何更有效地确定每个视频帧的输入分辨率是后续的一个重要的研究方向。可以考虑的工作包括利用强化学习、Gumbel Softmax采样等方式实现更加合理的自适应分帧策略。

7.3　未来可能的研究方向

本文针对行人重识别的时空建模进行深入和细致的研究，提出了时空交互和时序互补的完备特征学习方法，遮挡鲁棒的时空补全方法，以及一个高效时空建模框架。但视频行人重识别问题以及更广泛的重识别领域还有很多问题需要解决，而本文所研究的问题也需要进行更深入的研究和拓展。本节接下来将对行人重识别发展区域和未来研究方向阐述个人的一些见解。

（1）多模态信息的利用：目前行人重识别的研究大多集中在常规相机的可见光图像展开，只能解决光照充足下的重识别任务。在夜间或者光线很暗的场景，可见光摄像机无法得到清晰的彩色图像，此时需要利用红外摄像机在夜间拍摄到清晰的红外图像。但是，可见光图像和红外图像属于两种模态，捕捉的图像存在较大差异，因此已有成果无法直接应用到可见光图像和红外图像的跨模态行人重识别任务。为了满足夜间行人重识别应用的需求，需要设计相对的跨模态行人重识别算法 [126] 。在未来的研究中，可以考虑利用对抗方式 [127]

学习不同模态共享的特征，减小模态间的差异。

（2）领域适应行人重识别问题：当前行人重识别的训练集合测试集往往收集自相似的环境，研究方法通常在同一数据集的训练集训练并在该数据集的测试集上汇报相应的检测结果。近期的研究 [128] 发现，跨数据集的测试条件下，当前已有方法的行人重识别性能显著下降。为进一步提升模型在重识别任务上的泛化性能，需要设计相应的领域适应（Domain Adaptation）[129] 方法。在未来的研究中，可以考虑研究基于元学习 [130] 的方法，学习不同行人重识别任务之间的共性与特定，对跨域应用的重识别模型进行自动学习与调整。

（3）大规模无标注视频数据的使用：对于视频行人重识别而言，现阶段所获得的标注数据并不多，且标注成本很大。而与此同时，在监控摄像头上很容易采集大规模的无标注的行人视频。利用无标注的视频数据学习出完备且鲁棒的的行人特征是解决标注稀缺的一个解决思路。此外，视频中有很多先验知识可以利用，比如一个视频的所有帧都属于同一个行人。如何充分挖掘无标注视频中的先验知识，从而有效解决现阶段视频行人重识别标注数据不足的问题，构建更加鲁棒的重识别模型，也是未来的一个研究方向。

（4）基于视频的换衣行人重识别问题：目前，行人重识别被广泛应用于监控短时间内追踪行人，衣服颜色等特征是用于匹配的重要表观信息。当行人更换了衣服时，衣服颜色等表观信息也会失效，当前已有方法的行人重识别性能会显著下降。近年来研究者 [131] 开始尝试利用行人轮廓，深度图等其他模态信息辅助换衣重识别。但是轮廓和深度信息在一般的监控场景下让然难以获取。由于行人视频包含步态等动作信息，不受衣服变换的影响。因此，未来可以考虑学习从行人视频提取抗衣服变化的动作特征，构建更鲁棒的抗衣服变化的行人重识别模型。

参考文献

[1] Ye M, Shen J, Lin G, et al. Deep learning for person re-identification: A survey and outlook [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021.

[2] Zajdel W, Zivkovic Z, Krose B J. Keeping track of humans: Have i seen this person before? [C]//International Conference on Robotics and Automation. 2005: 2081-2086.

[3] Farenzena M, Bazzani L, Perina A, et al. Person re-identification by symmetry-driven accumulation of local features [C]//Computer Vision and Pattern Recognition. 2010: 2360-2367.

[4] Bazzani L, Cristani M, Perina A, et al. Multiple-shot person re-identification by hpe signature [C]//International Conference on Pattern Recognition. 2010: 1413-1416.

[5] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks [C]//Advances in Neural Information Processing Systems. 2012: 1097-1105.

[6] Yi D, Lei Z, Liao S, et al. Deep metric learning for person re-identification [C]//International Conference on Pattern Recognition. 2014: 34-39.

[7] Li W, Zhao R, Xiao T, et al. Deepreid: Deep filter pairing neural network for person re-identification [C]//Computer Vision and Pattern Recognition. 2014: 152-159.

[8] Dai Z, Chen M, Gu X, et al. Batch dropblock network for person re-identification and beyond [C]//International Conference on Computer Vision. 2019: 3691-3701.

[9] Xia B N, Gong Y, Zhang Y, et al. Second-order non-local attention networks for person re-identification [C]//International Conference on Computer Vision. 2019: 3760-3769.

[10] Zhang Z, Lan C, Zeng W, et al. Densely semantically aligned person re-identification [C]// Computer Vision and Pattern Recognition. 2019: 667-676.

[11] Zeng K, Ning M, Wang Y, et al. Hierarchical clustering with hard-batch triplet loss for person re-identification [C]//Computer Vision and Pattern Recognition. 2020: 13657-13665.

[12] Zhang Z, Lan C, Zeng W, et al. Relation-aware global attention for person re-identification [C]//Computer Vision and Pattern Recognition. 2020: 3186-3195.

[13] Zheng L, Zhang H, Sun S, et al. Person re-identification in the wild [C]//Computer Vision and Pattern Recognition. 2017: 1367-1376.

[14] Sun Y, Zheng L, Yang Y, et al. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline) [C]//European Conference on Computer Vision. 2018: 480-496.

[15] Sun Y, Zheng L, Deng W, et al. Svdnet for pedestrian retrieval [C]//International Conference on Computer Vision. 2017: 3800-3808.

[16] Varior R R, Haloi M, Wang G. Gated siamese convolutional neural network architecture for human re-identification [C]//European Conference on Computer Vision. 2016: 791-808.

[17] Varior R R, Shuai B, Lu J, et al. A siamese long short-term memory architecture for human re-identification [C]//European Conference on Computer Vision. 2016: 135-153.

[18] Wang Y, Chen Z, Wu F, et al. Person re-identification with cascaded pairwise convolutions [C]//Computer Vision and Pattern Recognition. 2018: 1470-1478.

[19] Cheng D, Gong Y, Zhou S, et al. Person re-identification by multi-channel parts-based cnn with improved triplet loss function [C]//Computer Vision and Pattern Recognition. 2016: 1335-1344.

[20] Hermans A, Beyer L, Leibe B. In defense of the triplet loss for person re-identification [J]. arXiv preprint arXiv:1703.07737, 2017.

[21] Chen W, Chen X, Zhang J, et al. Beyond triplet loss: a deep quadruplet network for person re-identification [C]//Computer Vision and Pattern Recognition. 2017: 403-412.

[22] Miao J, Wu Y, Liu P, et al. Pose-guided feature alignment for occluded person re-identification [C]//International Conference on Computer Vision. 2019: 542-551.

[23] Song C, Huang Y, Ouyang W, et al. Mask-guided contrastive attention model for person reidentification [C]//Computer Vision and Pattern Recognition. 2018: 1179-1188.

[24] Kalayeh M M, Basaran E, Gökmen M, et al. Human semantic parsing for person re-identification [C]//Computer Vision and Pattern Recognition. 2018: 1062-1071.

[25] Hou R, Ma B, Chang H, et al. Interaction-and-aggregation network for person re-identification [C]//Computer Vision and Pattern Recognition. 2019: 9317-9326.

[26] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C]//Advances in Neural Information Processing Systems. 2017: 5998-6008.

[27] He S, Luo H, Wang P, et al. Transreid: Transformer-based object re-identification [C]// International Conference on Computer Vision. 2021: 15013-15022.

[28] Ren M, He L, Liao X, et al. Learning instance-level spatial-temporal patterns for person re-identification [C]//International Conference on Computer Vision. 2021: 14930-14939.

[29] Zheng L, Shen L, Tian L, et al. Scalable person re-identification: A benchmark [C]// International Conference on Computer Vision. 2015: 1116-1124.

[30] Liang X, Gong K, Shen X, et al. Look into person: Joint body parsing & pose estimation network and a new benchmark [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 41(4): 871-885.

[31] Gong K, Liang X, Zhang D, et al. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing [C]//Computer Vision and Pattern Recognition. 2017: 932-940.

[32] Gao S, Wang J, Lu H, et al. Pose-guided visible part matching for occluded person reid [C]// Computer Vision and Pattern Recognition. 2020: 11744-11752.

[33] Wang G, Yang S, Liu H, et al. High-order information matters: Learning relation and topology for occluded person re-identification [C]//Computer Vision and Pattern Recognition. 2020: 6449-6458.

[34] He L, Wang Y, Liu W, et al. Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification [C]//International Conference on Computer Vision. 2019: 8450-8459.

[35] Zhao S, Gao C, Zhang J, et al. Do not disturb me: Person re-identification under the interference of other pedestrians [C]//European Conference on Computer Vision. 2020: 647-663.

[36] He L, Liu W. Guided saliency feature learning for person re-identification in crowded scenes [C]//European Conference on Computer Vision. 2020: 357-373.

[37] Chen P, Liu W, Dai P, et al. Occlude them all: Occlusion-aware attention network for occluded person re-id [C]//International Conference on Computer Vision. 2021: 11833-11842.

[38] Li Y, He J, Zhang T, et al. Diverse part discovery: Occluded person re-identification with part-aware transformer [C]//Computer Vision and Pattern Recognition. 2021: 2898-2907.

[39] Yang W, Huang H, Zhang Z, et al. Towards rich feature discovery with class activation maps augmentation for person re-identification. [C]//Computer Vision and Pattern Recognition. 2019: 1389-1398.

[40] Sun Y, Xu Q, Li Y, et al. Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification [C]//Computer Vision and Pattern Recognition. 2019: 393-402.

[41] Wei L, Zhang S, Gao W, et al. Person trasfer gan to bridge domain gap for person re-identification [C]//Computer Vision and Pattern Recognition. 2018: 79-88.

[42] Chen T, Ding S, Xie J, et al. Abd-net: Attentive but diverse person re-identification [C]// International Conference on Computer Vision. 2019: 8351-8361.

[43] Zheng L, Bie Z, Sun Y, et al. Mars: A video benchmark for large-scale person re-identification [C]//European Conference on Computer Vision. 2016: 868-884.

[44] Liu Y, Yan J, Ouyang W. Quality aware network for set to set recognition [C]//Computer Vision and Pattern Recognition. 2017: 4694-4703.

[45] Song G, Leng B, Liu Y, et al. Region-based quality estimation network for large-scale person re-identification [C]//AAAI Conference on Artificial Intelligence. 2018: 7347-7354.

[46] Li S, Bak S, Carr P, et al. Diversity regularized spatiotemporal attention for video-based person re-identification [C]//Computer Vision and Pattern Recognition. 2018: 369-378.

[47] Zhao Y, Shen X, Jin Z, et al. Attribute-driven feature disentangling and temporal aggregation for video person re-identification [C]//Computer Vision and Pattern Recognition. 2019: 4913-4922.

[48] McLaughlin N, del Rincon J M, Miller P C. Recurrent convolutional network for video-based person re-identification [C]//Computer Vision and Pattern Recognition. 2016: 1325-1334.

[49] Liao X, He L, Yang Z. Video-based person re-identification via 3d convolutional networks and non-local attention [C]//Asian Conference on Computer Vision. 2018: 620-634.

[50] Chen G, Rao Y, Lu J, et al. Temporal coherence or temporal motion: Which is more critical for video-based person re-identification? [C]//European Conference on Computer Vision. 2020: 660-676.

[51] Gu X, Chang H, Ma B, et al. Appearance-preserving 3d convolution for video-based person re-identification [C]//European Conference on Computer Vision. 2020: 228-243.

[52] Li J, Zhang S, Huang T. Multi-scale 3d convolution network for video based person re-identification [C]//AAAI Conference on Artificial Intelligence. 2019: 8618-8625.

[53] Yang J, Zheng W S, Yang Q, et al. Spatial-temporal graph convolutional network for video-based person re-identification [C]//Computer Vision and Pattern Recognition. 2020: 3289-3299.

[54] Liu X, Zhang P, Yu C, et al. Watching you: Global-guided reciprocal learning for video-based person re-identification [C]//Computer Vision and Pattern Recognition. 2021: 13334-13343.

[55] Liu H, Jie Z, Jayashree K, et al. Video-based person re-identification with accumulative motion context [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2017, 28(10): 2788-2802.

[56] Medsker L R, Jain L. Recurrent neural networks [J]. Design and Applications, 2001, 5: 64-67.

[57] Hochreiter S, Schmidhuber J. Long short-term memory [J]. Neural computation, 1997, 9(8): 1735-1780.

[58] Xu S, Cheng Y, Gu K, et al. Jointly attentive spatial-temporal pooling networks for video-based person re-identification [C]//International Conference on Computer Vision. 2017: 4743-4752.

[59] Zhou Z, Huang Y, Wang W, et al. See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification [C]//Computer Vision and Pattern Recognition. 2017: 6776-6785.

[60] Liu K, Ma B, Zhang W, et al. A spatiotemporal appearance representation for video-based pedestrian re-identification [C]//International Conference on Computer Vision. 2015: 3810-3818.

[61] Chen D, Li H, Xiao T, et al. Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding [C]//Computer Vision and Pattern Recognition. 2018: 1169-1178.

[62] Gao J, Nevatia R. Revisiting temporal modeling for video-based person reid [C]//British Machine Vision Conference. 2018.

[63] Zhang L, Shi Z, Zhou J T, et al. Ordered or orderless: A revisit for video based person re-identification [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 43(4): 1460-1466.

[64] Ji S, Xu W, Yang M, et al. 3d convolutional neural networks for human action recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 35(1): 221-231.

[65] Qiu Z, Yao T, Mei T. Learning spatio-temporal representation with pseudo-3d residual networks [C]//International Conference on Computer Vision. 2017: 5533-5541.

[66] Yan Y, Qin J, Chen J, et al. Learning multi-granular hypergraphs for video-based person re-identification [C]//Computer Vision and Pattern Recognition. 2020: 2899-2908.

[67] Aich A, Zheng M, Karanam S, et al. Spatio-temporal representation factorization for video-based person re-identification [C]//International Conference on Computer Vision. 2021: 152-162.

[68] Hou R, Ma B, Chang H, et al. Iaunet: Global context-aware feature learning for person reidentification [J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 32 (10): 4460-4474.

[69] He T, Jin X, Shen X, et al. Dense interaction learning for video-based person re-identification [C]//International Conference on Computer Vision. 2021: 1490-1501.

[70] Wang Y, Zhang P, Gao S, et al. Pyramid spatial-temporal aggregation for video-based person re-identification [C]//International Conference on Computer Vision. 2021: 12026-12035.

[71] Eom C, Lee G, Lee J, et al. Video-based person re-identification with spatial and temporal memory networks [C]//International Conference on Computer Vision. 2021: 12036-12045.

[72] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks [C]// International Conference on Learning Representations. 2016.

[73] Li W, Zhao R, Xiao T, et al. Deepreid: Deep filter pairing neural network for person re-identification. [C]//Computer Vision and Pattern Recognition. 2014: 152-159.

[74] Zheng Z, Zheng L, Yang Y. Unlabeled samples generated by gan improve the person re-identification baseline in vitro [C]//International Conference on Computer Vision. 2017: 3754-3762.

[75] Wang T, Gong S, Zhu X, et al. Person reidentification by video ranking [C]//European Conference on Computer Vision. 2014: 688-703.

[76] Wu Y, Lin Y, Dong X, et al. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning [C]//Computer Vision and Pattern Recognition. 2018: 5177-5186.

[77] Wei L, Zhang S, Yao H, et al. Glad: global-local-alignment descriptor for pedestrian retrieval [C]//Proceedings of ACM International Conference on Multimedia. 2017: 420-428.

[78] Hou R, Ma B, Chang H, et al. Feature completion for occluded person re-identification [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021.

[79] Felzenszwalb P F, Girshick R B, McAllester D, et al. Object detection with discriminatively trained part-based models [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 32(9): 1627-1645.

[80] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks [C]//Advances in Neural Information Processing Systems. 2015: 91-99.

[81] Office U H. i-lids multiple camera tracking scenario definition [J]. 2008.

[82] Dehghan A, Modiri Assari S, Shah M. Gmmcp tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking [C]//Computer Vision and Pattern Recognition. 2015: 4091-4099.

[83] Bolle R M, Connell J H, Pankanti S, et al. The relation between the roc curve and the cmc [C]// Fourth IEEE workshop on Automatic Identification Advanced Technologies. 2005: 15-20.

[84] Zheng L, Shen L, Tian L, et al. Scalable person re-identification: A benchmark [C]// International Conference on Computer Vision. 2015: 1116-1124.

[85] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition [C]//Computer Vision and Pattern Recognition. 2016: 770 - 778.

[86] Huang G, Liu Z, Weinberger K Q, et al. Densely connected convolutional networks [C]// Computer Vision and Pattern Recognition. 2016: 4700-4708.

[87] Singh K K, Lee Y J. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization [C]//International Conference on Computer Vision. 2017: 3544-3553.

[88] Zhang X, Wei Y, Feng J, et al. Adversarial complementary learning for weakly supervised object localization [C]//Computer Vision and Pattern Recognition. 2018: 1325-1334.

[89] Zheng M, Karanam S, Wu Z, et al. Re-identification with consistent attentive siamese networks [C]//Computer Vision and Pattern Recognition. 2019: 5735-5744.

[90] Suh Y, Wang J, Tang S, et al. Part-aligned bilinear representations for person re-identification. [C]//European Conference on Computer Vision. 2018: 402-419.

[91] Zhang S, Yang J, Schiele B. Occluded pedestrian detection through guided attention in cnns [C]//Computer Vision and Pattern Recognition. 2018: 6995 - 7003.

[92] Wang X, Girshick R, Gupta A, et al. Non-local neural networks [C]//Computer Vision and Pattern Recognition. 2018: 7794-7803.

[93] Hu J, Shen L, Sun G. Squeeze-and-excitation networks [C]//Computer Vision and Pattern Recognition. 2018: 7132-7141.

[94] Li J, Wang J, Tian Q, et al. Global-local temporal representations for video person re-identification [C]//International Conference on Computer Vision. 2019: 3958-3967.

[95] Kingma D P, Ba J. Adam: A method for stochastic optimization [J]. arXiv preprint arXiv:1412.6980, 2014.

[96] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions [C]//Computer Vision and Pattern Recognition. 2015: 1-9.

[97] Subramaniam A, Nambiar A, Mittal A. Co-segmentation inspired attention networks for video-based person re-identification [C]//International Conference on Computer Vision. 2019: 562-572.

[98] Su C, Li J, Zhang S, et al. Pose-driven deep convolutional model for person re-identification [C]//International Conference on Computer Vision. 2017: 3960-3969.

[99] Chang X, Hospedales T M, Xiang T. Multi-level factorisation net for person re-identification [C]//Computer Vision and Pattern Recognition. 2018: 2109-2118.

[100] Shen Y, Xiao T, Li H, et al. End-to-end deep kronecker-product matching for person re-identification [C]//Computer Vision and Pattern Recognition. 2018: 6886-6895.

[101] Wang C, Zhang Q, Huang C, et al. Mancs: A multi-task attentional network with curriculum sampling for person re-identification [C]//European Conference on Computer Vision. 2018: 365-381.

[102] Chen D, Xu D, Li H, et al. Group consistent similarity learning via deep crf for person re-identification [C]//Computer Vision and Pattern Recognition. 2018: 8649-8658.

[103] Sarfraz M S, Schumann A, Eberle A, et al. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking [C]//Computer Vision and Pattern Recognition. 2018: 420-429.

[104] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module [C]//European Conference on Computer Vision. 2018: 3-19.

[105] Ghiasi G, Lin T Y, Le Q V. Dropblock: A regularization method for convolutional networks [C]//Advances in Neural Information Processing Systems. 2018: 10727-10737.

[106] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift [C]//International Conference on Machine Learning. 2015: 448-456.

[107] Zhong Z, Zheng L, Kang G, et al. Random erasing data augmentation [C]//AAAI Conference on Artificial Intelligence. 2020: 13001-13008.

[108] van der Maaten L, Hinton G. Visualizing data using t-sne [J]. Journal of Machine Learning Research, 2008, 9(86): 2579-2605.

[109] Zhuo J, Chen Z, Lai J, et al. Occluded person re-identification [C]//International Conference on Multimedia and Expo. 2018: 1-6.

[110] Huang H, Li D, Zhang Z, et al. Adversarially occluded samples for person re-identification [C]//Computer Vision and Pattern Recognition. 2018: 5098-5107.

[111] He L, Liang J, Li H, et al. Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach [C]//Computer Vision and Pattern Recognition. 2018: 7073-7082.

[112] Pathak D, Krahenbuhl P, Donahue J, et al. Context encoders: Feature learning by inpainting [C]//Computer Vision and Pattern Recognition. 2016: 2536-2544.

[113] Albawi S, Mohammed T A, Al-Zawi S. Understanding of a convolutional neural network [C]//International Conference on Engineering and Technology. 2017: 1-6.

[114] Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions [J]. arXiv preprint arXiv:1511.07122, 2015.

[115] Clevert D A, Unterthiner T, Hochreiter S. Fast and accurate deep network learning by exponential linear units (elus) [J]. arXiv preprint arXiv:1511.07289, 2015.

[116] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation [C]//International Conference on Medical image computing and computer-assisted intervention. 2015: 234-241.

[117] Isola P, Zhu J Y, Zhou T, et al. Image-to-image translation with conditional adversarial networks [C]//Computer Vision and Pattern Recognition. 2017: 1125-1134.

[118] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets [C]//2014.

[119] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks [J]. arXiv preprint arXiv:1511.06434, 2015.

[120] Hou R, Chang H, Ma B, et al. Temporal complementary learning for video person re-identification [C]//European conference on computer vision. 2020: 388-405.

[121] Hou R, Ma B, Chang H, et al. Vrstc: Occlusion-free video person re-identification [C]// Computer Vision and Pattern Recognition. 2019: 7183-7192.

[122] Zhao L, Li X, Wang J, et al. Deeply-learned part-aligned representations for person re-identification [C]//International Conference on Computer Vision. 2017: 3239 - 3248.

[123] He L, Sun Z, Zhu Y, et al. Recognizing partial biometric patterns. [J]. arXiv preprint arXiv:1810.07399, 2018.

[124] Li W, Zhu X, Gong S. Harmonious attention network for person re-identification [C]// Computer Vision and Pattern Recognition. 2018: 2285 - 2294.

[125] Ge Y, Li Z, Zhao H, et al. Fd-gan: Pose-guided feature distilling gan for robust person re-identification. [C]//Advances in Neural Information Processing Systems. 2018: 1222-1233.

[126] Li D, Wei X, Hong X, et al. Infrared-visible cross-modal person re-identification with an x modality [C]//AAAI Conference on Artificial Intelligence. 2020: 4610-4617.

[127] Creswell A, White T, Dumoulin V, et al. Generative adversarial networks: An overview [J]. IEEE Signal Processing Magazine, 2018, 35(1): 53-65.

[128] Ge Y, Chen D, Li H. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification [C]//International Conference on Learning Representations. 2020.

[129] Wang M, Deng W. Deep visual domain adaptation: A survey [J]. Neurocomputing, 2018, 312: 135-153.

[130] Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks [C]//International Conference on Machine learning. 2017: 1126-1135.

[131] Wu A, Zheng W S, Lai J H. Robust depth-based person re-identification [J]. IEEE Transactions on Image Processing, 2017, 26(6): 2588-2603.

致 谢

时光飞逝，六年的博士生涯转瞬即逝。值此博士论文即将完成之际，本人衷心感谢所有给予我关心、帮助、支持的老师、同学、朋友以及家人们。

首先要感谢我的导师常虹研究员。常老师是我进入人工智能研究领域的启蒙老师。从大四到计算所实习开始，常老师便全程参与了我每个课题的研究，从问题选择、方法设计、文章写作每一个步骤，常老师都给予了我充分的指导，甚至不惜牺牲休息时间为我修改文章和讨论问题。常老师的详尽指导帮助我提升独立思考和解决问题的能力，也建立了良好的科研品味和科学素养。我过去几年的成长离不开常老师的细致指导，我取得的每一次进步也都包含了常老师的指导和付出。从常老师身上我感受到了一个研究者对于科研的热爱和追求，未来的科研道路我将秉承这份执着和专注。很荣幸能够成为常老师的学生，师恩难忘，我会永远铭记常老师一直以来给予我的关心、鼓励、指导和帮助。

感谢课题组的马丙鹏老师。马老师在博士论文的选题上给予了我详尽的帮助，引领我进入行人重识别这一研究课题。回想起最初接触行人重识别课题的学业生涯，每周都要和马老师多次讨论，最终促成了第一个研究工作的产出。每次遇到问题，马老师总是跟我进行深入细致讨论，帮助我理清思路。在论文写作上，马老师也耐心指导如何科学组织论文结构，让我获益良多。感谢马老师多年来的指导和帮助，也感谢马老师对我宽容和信任。

然后要感谢的是山世光研究员。过去的几年里，山老师为我的研究课题提供了宝贵的指导意见，同时开拓了我的研究思路。尽管山老师身兼多职，但他却总会在百忙之后抽出时间与我讨论，帮助我修改文章，以及解决科研和工作的问题。特别是，在博士最后一年找工作期间，山老师主动帮我联系了多个知名课题组，让我坚定了投身科研的决心。在最后的留所答辩中，从答辩报告最初的撰写、中间十几次的修改、到最终的演讲展示，山老师都全程参与其中，为我提供了大量的指导，让我受益良多。十分感谢山老师一直以来对我科研和工作的关心、指导和帮助，希望以后能有更多机会向山老师学习，也希望自己不负山老师的信任和帮助。

感谢实验室主任陈熙霖老师。陈老师为我们提供了一流的科研坏境和良好

的学术氛围。陈老师思维缜密、想法独到。在与陈老师的多次学术讨论中，陈老师对研究问题的见解，对我研究工作的批判，总会引导我重新深入思考自己的研究领域和研究工作，帮助我提升了独立思考的能力。陈老师对于科研的严谨和坚定帮助我提升了自己的科研素养，激励我以更高的科研水准要求自己。

感谢实验室其他老师对我的指导和帮助，他们是王瑞平老师、阚美娜老师、韩琥老师、杨双老师、曾加贝老师、张杰老师、许倩倩老师、李亮老师、王树徽老师、蒋树强老师、黄庆明老师，与各位老师的交流让我受益匪浅。感谢实验室辛勤工作的王骐老师、秘书王晓彪老师、胡兰平老师、程一老师、班瀚文老师，为我们的科研工作提供了强大的后勤保障，让我可以一心投入科研，不必为琐事操心。感谢研究生部的周世佳老师、张平老师、李丹老师在我生活、申请奖励、答辩、就业方面予以的热心帮助。

感谢在计算所求学期间里，邬书哲师兄、李勇师兄、徐霞清师姐、梁孔明师兄、尹肖贻师兄、李煜林师兄、李燊师兄、何振梁师兄、胡兰青师姐、刘志鹏师兄、姜华杰师兄、刘昊淼师兄以及其他所有帮助过我的师兄师姐们。他们在我的科研以及生活中给予了我很多的帮助。感谢小组同学谷欣乾、张弘楷、陈玉成、包立强、徐富荣、毛凤玲、王泽宇、王诚、周林茂、张翔州、康楠、柏淑涛、郭纵、张苡卉、崔晏菲、胡民阳、李胤祺、叶博涛、伍子强、沈鸿泽、刘宇奇、赵家和、梁佳宸以及其他所有小组成员。正是这个温暖的集体，为我的博士生涯提供了一个美好的回忆。感谢经常互相交流学习的韩春瑞、刘雪静、牛雪松、何晨、毛懿荣、骆石英、赵越、王玺珺、闵越聪、李坤彦、申震、赖楠、黎向阳、卓君宝、胡梦颖、王文彬、杨志勇、姜洋邦彦、孙蕴嘉。他们让我的博士生活充满了欢声笑语。祝所有的同学都能够以梦为马，不负韶华。

感谢我的家人。感谢我的父母总是在远方牵挂着我，给予我物质和精神的支持。特别是感谢我的母亲对我一直以来深切的爱和无条件的包容。感谢我的老公潘虹宇先生，我们在计算所相遇、相识、相知、相爱、相许。感谢潘先生在我读博期间对我在科研和生活上的帮助和支持，在我每次迷茫的时候开导我，在科研不顺的时候帮我分析问题，跟我一起熬夜修改论文等等。

最后，祝愿所有的老师、同学、朋友以及家人们身体健康。

作者简历及攻读学位期间发表的学术论文与研究成果

作者简历：

侯瑞兵，河北省邢台市人，中国科学院计算技术研究所博士研究生。

教育经历：

2016 年 9 月— 至今，中国科学院计算技术研究所，计算机应用技术专业，博士生

2012 年 9 月— 2016 年 7 月，西北工业大学，计算机科学与技术专业，本科生

已发表的学术论文：

1. **Ruibing Hou**, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, Xilin Chen. Feature Completion for Occluded Person Re-Identification [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI), 2021.
2. **Ruibing Hou**, Hong Chang, Bingpeng Ma, Rui Huang, Shiguang Shan. Bicnet-tks: Learning efficient spatial-temporal representation for video person re-identification [C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021.
3. **Ruibing Hou**, Hong Chang, Bingpeng Ma, Shiguang Shan, Xilin Chen. Temporal complementary learning for video person re-identification [C]//European Conference on Computer Vision (ECCV), 2020.
4. **Ruibing Hou**, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, Xilin Chen. IAUnet: Global context-aware feature learning for person reidentification [J]//IEEE Transactions on Neural Networks and Learning Systems (TNNLS), 2020, 32(10), 4460-4474.
5. **Ruibing Hou**, Hong Chang, Bingpeng Ma, Shiguang Shan, Xilin Chen. Cross attention network for few-shot classification [C]//Advances in Neural Information Processing Systems (NeurIPS), 2019.

6. **Ruibing Hou**, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, Xilin Chen. Interaction-and-aggregation network for person re-identification [C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
7. **Ruibing Hou**, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, Xilin Chen. VRSTC: Occlusion-free video person re-identification [C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
8. **Ruibing Hou**, Hong Chang, Bingpeng Ma, Xilin Chen. Video prediction with bidirectional constraint network [C]//IEEE International Conference on Face and Gesture Recognition (FG), 2019.

攻读博士学位期间已参加的科研项目

- 国家自然科学基金面上项目：面向真实监控场景的行人重识别关键问题研究，2019 年 1 月-2022 年 12 月。
- 国家自然科学基金面上项目：面向视觉特征表示的元学习方法研究，2020 年 1 月-2023 年 12 月。
- 北京市科委科技计划专项：基于视觉计算的道路交通异常智能识别技术研究与应用示范，2018 年 12 月-2020 年 5 月。

攻读博士学位期间的获奖情况

- 中科院所长特别奖（夏培肃奖），2021
- 中国科学院大学三好学生标兵，2020
- 博士生国家奖学金，2019
- 中国科学院大学三好学生，2019