



上海交通大学学位论文

基于深度学习的行人 ReID

姓 名: KEN YUEN TAN

学 号: 519030990031

导 师: 周越副教授

学 院: 电子信息与电气工程学院

学科/专业名称: 自动化

申请学位层次: 学士

2023 年 05 月

A Dissertation Submitted to
Shanghai Jiao Tong University for Bachelor Degree

DEEP LEARNING-BASED PERSON
RE-IDENTIFICATION


Author: Ken Yuen Tan
Supervisor: Associate Prof. Zhou Yue.

School of Electronic Information and Electrical Engineering
Shanghai Jiao Tong University
Shanghai, P.R.China
June 12th, 2023

上海交通大学

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。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作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全知晓本声明的法律后果由本人承担。

学位论文作者签名：

日期：2023 年 5 月 21 日

上海交通大学

学位论文使用授权书

本人同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。

本学位论文属于：

☒公开论文

☐内部论文，保密☐1 年/☐2 年/☐3 年，过保密期后适用本授权书。

☐秘密论文，保密____年（不超过 10 年），过保密期后适用本授权书。

☐机密论文，保密____年（不超过 20 年），过保密期后适用本授权书。

（请在以上方框内选择打“√”）

学位论文作者签名：



日期：2023 年 5 月 21 日

指导教师签名：

日期：2023 年 5 月 21 日

摘 要

行人重识别任务是近期计算机视觉领域的热门任务之一，其主要目的是在多个不同、无重叠视角的监控摄像头之间准确识别同一行人。该任务具有广泛的实际应用场景，如视频监控、安全检测和人群管理等领域。

然而，行人重识别面临诸多挑战，如遮挡、光照、视角变化、背景复杂等问题。为了解决这些问题，基于视觉 Transformer 的行人重识别方法由于相比 CNN 有更强的序列信息捕捉能力，并保留图像完整信息，而得到广泛研究和应用。然而，目前视觉 Transformer 仍然存在计算复杂度高，效率较低的问题。

此外，单靠全局和局部特征无法完全区分两个行人的差别，因此近年来行人重识别任务的研究方向逐渐往辅助信息的结合发展。然而，采集和整理辅助信息非常复杂，增加网络复杂度。依靠现有特征图进行更为细致的数据增强和特征融合，更好利用已有数据和特征信息是重要研究方向之一。

为了解决这些问题，本文提出基于数据增强和特征融合重构的行人重识别网络，该网络包括四个模块。其中，基于数据增强的行人重识别网络通过三个数据增强模块有效降低噪音，提升特征表示能力，增强局部特征学习。同时，基于特征融合与重构的行人重识别网络增强输出特征块的空间相关性。

本文提出的行人重识别网络具有很好的应用前景。与传统的卷积神经网络相比，该网络通过数据增强和特征融合重构等方法有效地提高了行人重识别任务的性能和效率。

关键词：行人重识别，Transformer，数据增强，特征融合

ABSTRACT

Person re-identification(ReID) is one of the hot research topics in computer vision, which aims to accurately identify the same person across multiple non-overlapping surveillance cameras with different viewpoints. This task has wide real world applications such as video surveillance, security detection, and crowd management.

However, person ReID faces many challenges, such as occlusions, illumination variations, viewpoint changes, complex backgrounds and etc. In recent years, deep learning-based methods have been widely studied and applied to address these challenges. Among them, vision transformers have stronger sequence information capturing ability and retain image integrity compared to CNNs, thus vision transformer models have been widely used. However, vision transformers still suffer from high computational complexity and low efficiency.

In addition, as global and local features alone cannot fully differentiate the differences between two persons, person ReID researchers has gradually shifted the direction towards the integration of auxiliary information. However, collecting and organizing auxiliary information is complex and increases the complexity of the network. Therefore, relying on existing feature maps for more detailed data augmentation and feature fusion, as well as better utilizing existing data and feature information, are significant research directions.

To address these issues, this paper proposes a person ReID network based on data augmentation and feature fusion & reconstruction, which includes four modules. Among them, the data augmentation-based person ReID network effectively reduces noise, enhances feature representation ability, and strengthens local feature learning through three data augmentation modules, thereby improving the model's performance. At the same time, the feature fusion and reconstruction-based person ReID network enhances the spatial correlation of the output feature block, reduces noise impact, and improves model performance.

The proposed person ReID network in this paper has good practicality and application prospects. Compared with traditional convolutional neural networks, this network

effectively improves the performance and efficiency of person ReID tasks through methods such as data augmentation and feature fusion reconstruction.

Key words: Person re-identification, Transformer, Data Augmentation, Feature Fusion

目 录

摘 要 I

ABSTRACT II

第一章 绪论 1

 1.1 研究背景与意义 1

 1.2 本文研究主要内容 2

 1.3 本章小结 4

第二章 相关背景知识介绍 5

 2.1 行人重识别任务 5

 2.1.1 行人重识别数据集 5

 2.1.2 表征学习 7

 2.1.3 度量学习 8

 2.2 TRANSFORMER 9

 2.2.1 自注意力机制原理 9

 2.2.2 Transformer 网络架构 10

 2.3 视觉 TRANSFORMER 11

 2.4 TRANSREID 网络 12

 2.4 本章小结 14

第三章 基于数据增强与特征融合的行人重识别网络 15

 3.1 网络架构 15

 3.1.1 行人重识别任务主干网络的选择 16

 3.1.2 网络新增模块 16

 3.2 输入图像块全尺寸增强模块 16

 3.2.1 具体实现方法 16

3.2.2 实验数据	17
3.3 输入图像块随机屏蔽模块	19
3.3.1 具体实现方法	19
3.3.2 现有方法对比	20
3.3.3 实验数据	20
3.4 输出特征批次方块随机屏蔽模块	24
3.4.1 具体实现方法	25
3.4.2 现有方法对比	26
3.4.3 实验数据	27
3.5 特征融合与重构模块	29
3.5.1 具体实现方法	29
3.5.2 实验数据	30
3.6 本章小结	32
第四章 实验结果与分析	33
4.1 图表格式实验环境及设置	33
4.1.1 网络训练环境与配置	33
4.1.2 网络参数	33
4.1.3 行人重识别数据集概述	33
4.1.4 网络性能评价指标	34
4.2 实验结果及分析	35
4.3 消融实验	36
4.3.1 网络架构研究	36
4.3.2 输出特征随机屏蔽模块	36
4.4 本章小结	38
第五章 结论	39
5.1 本文主要工作	39
5.1.1 基于数据增强的行人重识别网络	39

5.1.2 基于特征融合与重构的行人重识别网络 39

5.2 非技术性分析 40

5.2.1 法律分析 40

5.2.2 公共安全分析 40

5.2.3 对环境与社会的影响 40

5.3 未来研究工作展望 41

参 考 文 献.....43

致 谢47

第一章 绪论

1.1 研究背景与意义

行人重识别任务(Person ReIdentification, Person ReID)最早于 1996 年提出,并在 2006 年首次在 CVPR 会议中正式提出其概念。行人重识别任务的提出与为了公共安全而大规模建立的公共场合监控系统有关。监控系统可以通过分析各个区域的人流量进行相对应的调整,并起到防止犯罪活动的效果。尽管如此,依靠人类对监控系统采集的海量数据中追踪行人的移动轨迹非常困难,因此行人重识别技术应运而生。行人重识别网络不仅让警方通过行人移动轨迹快速锁定嫌疑犯的去向,对打击犯罪的效果有更大帮助,也可以寻找失踪人士。此外,此技术不仅可用于无人超市和智能机器人领域,减低企业对人力的依赖,也可以用于商场。此技术可以分析顾客在商场的运动轨迹和区域停留时间,追踪顾客的购物行为,了解顾客偏好,从而提高销售额。

目前行人重识别任务经历了从度量学习到深度学习的发展。随着视觉 Transformer 的出现,越来越多行人重识别任务开始使用视觉 Transformer 作为主干网络,或将 CNN 和视觉 Transformer 进行融合,以提升网络性能。尽管视觉 Transformer 相较 CNN 有诸多好处,但其占用的训练时间和计算与显存开销较多。目前研究人员提出视觉 Transformer 的各种变体,如 AAformer^[1]、SKIPAT^[2]、OH-Former^[3]、Per-former^[4]、DC-Former^[5]和 PAT^[6],从不同角度解决目前视觉 Transformer 的问题,并符合行人重识别任务的要求。尽管如此,更换行人重识别网络的主干网络工作量巨大,且这些视觉 Transformer 的内部网络架构都有一定程度的更动,导致应用场景的缩小。因此研究即插即用、相对简单且能融合现有行人重识别主干网络以解决现有问题的模块具有重大意义。

此外,仅仅依靠全局特征和局部特征并无法完全分辨两个行人的差别,因此近期行人重识别任务研究方向转为使用行人的辅助信息以提高性能。如 AD-ViT^[7]通过另一个行人属性识别网络预测行人图像的属性,并使用预测得来的属性作为行人重识别网络的输入。尽管效果优秀,行人辅助信息的采集与整理让行人重识别网络愈发复杂。TransReID^[8]和 PFT^[9]为行人重识别任务提供不同思路,即使不依靠额外的复杂辅助信息与辅助信息处理网络,仅依靠现有特征图进行更为细致的数据增强和进一步的特征融合也可以达成相同或甚至更强的性能,因此这方面的研究也具有积极意义。

从学术层面来说,不牺牲过多性能的前提下缩短训练时间、减少计算和显存开销可以有效提高网络效率。此外,训练时间过长可能导致网络记住训练集中的噪声和细节,因此缩短训练时间可以使网络更快进行迭代并收敛到最佳状态,减少过拟合的风险。此外,缩短训练时间可以让网络更快适应新的数据和环境,提高网络实时性能。当数据集变得更大或网络架构变得更复杂时,训练深度学习网络的计算开销通常也会增加。减少计算开销可以使网络更容易扩展到更大的数据集和更复杂的网络架构,从而提高网络性能和应用范围。

从现实应用层面来说,减少计算和显存开销可以让网络运行在计算资源有限的设备或环境,如嵌入式设备或移动设备。这些设备具有低成本和低功耗的特点,因此减少这些开销可以有效降低硬件成本,从而提高部署的成本效益。即使资源充足,大型企业部署的深度学习网络需要消耗大量能源,因此能源成本的下降对环境可持续性非常重要。此外,无需额外使用辅助信息和增加数据集大小对降低部署行人重识别网络的难度与成本有一定帮助。

从宏观层面来看,人工智能浪潮势不可挡,各个领域都尝试通过使用人工智能减低人力成本,提升企业运作效率。尽管如此,目前过于巨大的人工智能训练开销导致只有少数大型企业可以自行训练人工智能并投入应用,长期下来有可能导致少数几间企业掌握能颠覆人类社会的人工智能,形成人工智能霸权。降低训练人工智能的门槛,提升人工智能效率能让大量中小型企业能根据自身要求训练合适的人工智能,对真正做到人工智能大规模普及和应用具有积极意义。

1.2 本文研究主要内容

本文的研究内容基于行人重识别任务背景,致力探讨基于深度学习的行人重识别技术。本文将采用以视觉 Transformer 为主干网络的 TransReID,并在此基线网络上展开研究,旨在提高网络效率、增强局部特征学习,并加强输出特征序列之间的空间相关性。具体的研究内容包括:

1. 基于数据增强的行人重识别网络

承接上文,解决视觉 Transformer 的效率问题是其中重点研究内容之一。本文将融合 TransReID 网络和输入图像块序列随机屏蔽模块达成上述目标,通过输入图像块序列中的图像块进行随机屏蔽,减少视觉 Transformer 的输入以提升网络效率,减少训练时间、计算和显存开销。此外,此模块也有数据集正则化的作用。

行人重识别任务的主要挑战包括遮挡物或行人遮挡、光照、视角变化、行人穿搭变化、行人姿势、室内外环境变化、相机分辨率等因素。这些因素导致行人判断信息减少，且背景噪声更多，因此无法准确提取与行人相关的特征表示，进而降低识别成功率。行人重识别网络必须考虑上述所有因素，提高网络鲁棒性与泛化能力，以更好适应现实环境变化。然而，目前行人重识别数据集大小并不足以涵盖上述所有因素，网络也更容易过拟合。面对这个问题，目前业界也提出多个数据增强模块，如 Cutout^[10]、PatchSwap^[11]、FenceMask^[12]、Mixup^[13]、Hide-and-Seek^[14]等。数据增强有效缓解行人重识别数据不足的困境，并达成数据集正则化与避免网络过拟合的目的。



图 1-1 行人重识别主要挑战 (a) 裁切失误 (b) 图像模糊
(c) 遮挡 (d) 光照变化 (e) 穿搭相似 (f) 行人穿搭变化^[18]

除了融合上述输入预处理模块，输出特征序列的随机屏蔽模块也有助于网络提取强鲁棒性的特征。由于脸、手、脚等身体部位会随着视角的变化而变得不稳定，因此网络会集中在身体主要部位中提取特征，而忽略其他具有描述性的身体部位，从而降低网络处理遮挡行人数据集的有效性。因此，提取局部区域特征成为其中一个研究对象。本文提出适用于视觉 Transformer 的输出特征序列批次随机擦除模块，在同一批

次的输出特征序列随机擦除相同位置的特征序列块。这样相等于删除相同语义的身体部位，让网络专注提取被删除特征之外的局部区域特征。

2. 基于特征融合与重构的行人重识别网络

视觉 Transformer 将完整图像分割成输入图像块序列，但输入图像块之间仍然有一定程度上的关联，而经过 ViT 网络的特征后仍然存在这种关联性。网络可以分析并利用输出特征序列块空间相关性，以提取更具辨识度的特征。寻找输出特征序列块的空间相关性成为基于视觉 Transformer 行人重识别任务的研究对象之一。本文提出输出特征序列融合与重构模块，对输出特征序列进行分割与融合，最终使用经重构的输出特征序列作为分类器输入，以放大部分特征的重要性，提升网络性能。

1.3 本章小结

本章初步介绍行人重识别任务的研究背景和其应用，并说明行人重识别网络目前遇到的问题。针对这些问题，研究人员持续对行人重识别任务进行研究对学术和现实都具有积极意义。之后，本章阐述本文主要的研究内容，即通过数据增强与特征融合的方式提升网络效率、增强局部特征学习和增强输出特征序列之间的空间相关性，由此避免使用复杂的辅助信息，不依靠其他外来的辅助信息提升网络性能。

第二章 相关背景知识介绍

2.1 行人重识别任务

行人重识别，又称行人再识别，为当前计算机视觉领域热门任务之一，其主要任务为从多个不同、无重叠视角摄像头中捕捉行人图像，并通过计算机视觉技术提取行人特征，如衣服、体态、姿势等，并以此识别和检索行人身份。行人重识别技术主要应用在客流量较大的大型公共场所监控系统，如火车站、购物中心、机场和医院等。除了行人重识别技术，识别行人身份也可以依靠成熟的人脸识别技术，但监控系统不同的摄像头分辨率和拍摄角度导致监控系统无法获取行人正脸特征，而是更多获取行人的侧脸和后脑勺。其次，即使取得行人正脸，照片不高的解析度导致网络难以做出准确判断，因此人脸识别技术不适用于上述场景。行人重识别任务主要有两个实现方式，即表征学习和度量学习。

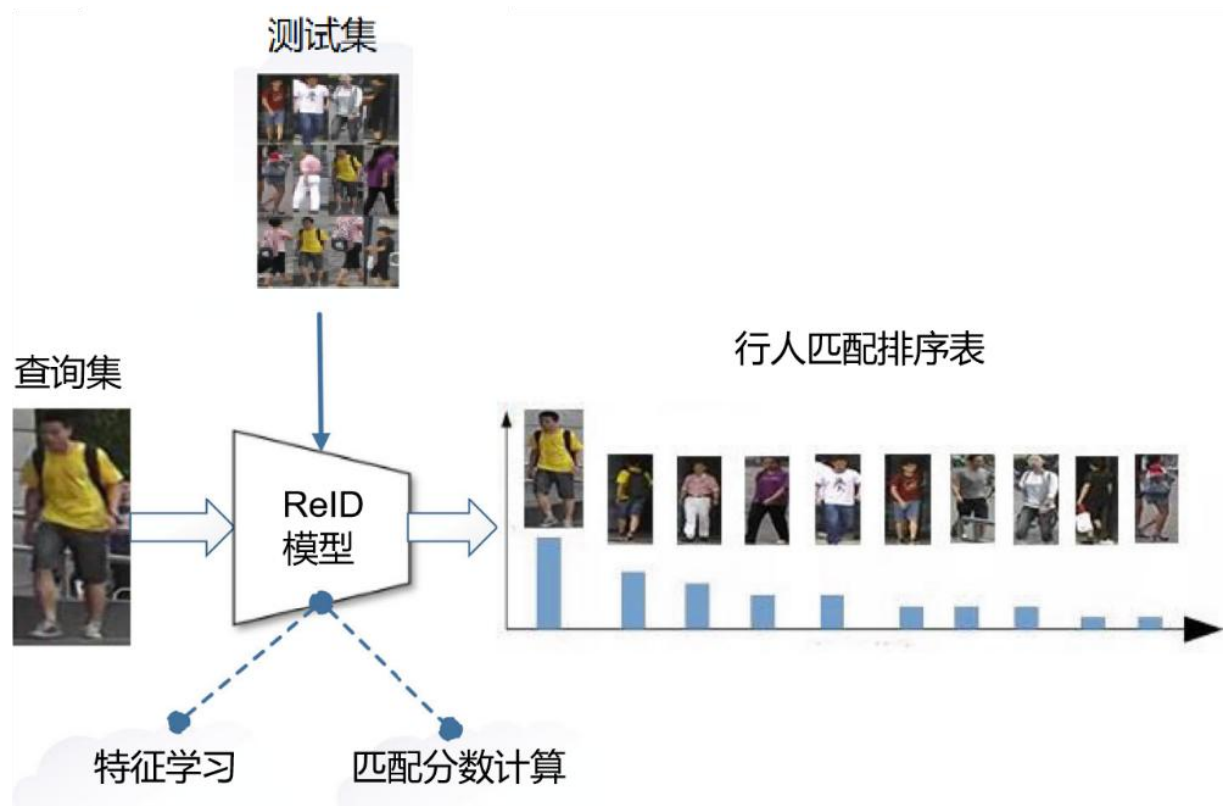


图 2-1 行人重识别网络示意图^[16]

2.1.1 行人重识别数据集

行人重识别数据集分为图像数据集和视频数据集，不同数据集对应不同的行人重识别任务研究方法。视频数据集的规模比图像数据集大得多，因为视频数据包含时间维度，每个行人在不同的时间点都可能出现在不同的位置，因此需要采集更多的数据来训练网络。此外，视频数据集则需要网络在整个视频序列上提取特征，这可能需要更复杂的网络架构和更大的计算资源。

表 2-1 热门行人重识别数据集参数

数据集	年份	总行人数量	总图像数量	摄像头数量
CUHK03 ^[17]	2014	1360	13164	10
Market-1501 ^[18]	2015	1501	32668	6
Airport ^[19]	2017	9651	39902	6
DukeMTMC-reID ^[20]	2017	1404	36411	8
MSMT17 ^[21]	2018	4101	126441	15

表 2-2 热门遮挡行人重识别数据集参数

数据集	年份	总行人数量	总图像数量
Occluded REID ^[22]	2018	200	2000
Partial-REID ^[23]	2015	60	900
iLIDS-VID ^[24]	2018	300	600(图像序列)
Occluded Duke ^[25]	2019	1404	35489

训练有效、泛化能力与鲁棒性强的行人重识别网络的前提为拥有总行人数量和总图像数量庞大的行人重识别数据集，其中数据集包括行人边界框和训练数据标注。相较于一般物体识别数据集百万数量级的数据，大部分行人重识别数据集的总图像数量仅仅是以万为计算单位，因此行人重识别网络更容易过拟合。此外，研究人员需要在同一地点同时使用多个摄像机采集行人重识别数据，对采集环境要求较高。最后，大规模采集行人重识别数据有侵犯隐私之嫌，因此行人重识别数据集的整体采集难度较

大。此外,为了让行人重识别网络更贴近现实环境,遮挡行人重识别数据集的采集具有重要意义。

一般上,行人重识别数据集分为训练集、测试集和查询集。网络使用训练集进行训练后,网络将会提取查询集与测试集图像的特征,并计算两者之间的相似度,对于每个查询图像中从测试集中找出前 N 个与其相似的图像。

2.1.2 表征学习

表征学习主要分为四大类,分别是全局表征学习、局部表征学习、辅助表征学习和视频表征学习,其目的为学习能够应对不同摄像头下行人变化的特征。^[26]

全局表征学习是一种从行人图像中提取全局特征向量的技术。由于深度神经网络被应用在图像分类问题,因此此方向为早期深度学习技术融合行人重识别任务的最初选择。全局特征学习 IDE 网络^[27]通过将每个行人视为不同类别,将行人重识别问题视作多类分类问题。此思路最终被研究人员广泛接受并使用。为了增强全局表征学习的效果和鲁棒性,研究人员也广泛探索了一些改进思路,如注意力机制^[28, 29]。

局部表征学习可以针对不同身体部位或区域分别进行表征学习,并聚合这些特征,让网络对身体部分错位具有鲁棒性。对图像中各个身体部位的识别成为局部表征学习的重要一环。一般上,身体部位识别可通过人体解析或姿势估计网络^[30, 31],和对图像进行切割^[32]。人体解析或姿势估计网络虽然可以提取对齐身体部位特征,但需要额外增加姿势检测器。切割法更为灵活,但是对遮挡和杂乱背景等杂讯敏感。

辅助表征学习通过为网络增加额外辅助信息,提升网络对不同行人的判断能力。这些辅助信息可以是域信息^[33](如每一个摄像机下的数据表示一类域)、语义信息^[34](如行人属性、衣着、性别等)、GAN 生成的信息^[35]等。辅助表征学习需要研究人员采集更多信息,其中一种方式是为原先没有这些信息的数据集进行人工标注,耗时巨大,而另一种方法为使用神经网络预测数据集的辅助信息,并假定这些预测值为真值。

视频表征学习是近期热门的行人重识别任务方向,其中每个行人都由多个帧的视频序列表示,并从中提取行人特征。与其他方式相比,数据含有时间信息和更为丰富的细节,且分析视频序列中的多个帧更加符合现实环境。尽管如此,视频表征学习需要面对提取时间信息^[36]、视频序列长短不一^[37]、视频序列中的异常值帧^[38]等一系列挑战。

随着对行人重识别任务的深入研究,目前行人重识别网路并不会局限于使用其中

一个表征学习方式,如本文参考的 TransReID 同时使用了全局表征学习、局部表征学习和辅助表征学习,将图像分割成多个图像块,并结合各个辅助信息作为主干网络输入进行网络训练。

2.1.3 度量学习

在应用深度学习处理行人重识别任务之前,度量学习是行人重识别任务的主要研究方向。度量学习将提取的行人特征映射到新的空间,使相同的人相距更近,不同的人相距更远。目前不同的损失函数已取代度量学习,引导表征学习网络提取更加准确的行人特征。目前基于深度学习的行人重识别任务常用损失函数包括身份损失(Identity Loss)、验证损失(Verification Loss)和三元损失(Triplet Loss)。不同的行人重识别任务研究角度对应不同损失函数的使用。此外,多个不同的损失函数可以相互结合使用,以增强网络性能。以下是其中三个不同行人重识别任务研究角度所对应的损失函数:

1. 图像分类问题:同一个行人的不同图片当成一个类别,并对行人照片进行分类。

身份损失通过交叉熵计算,其公式为

$$\mathcal{L}_{ID} = -\frac{1}{n} \sum_{i=1}^n \log(p(y_i|x_i)) \quad (2-1)$$

其中 n 为每个 batch 训练的样本数, x_i 为每个 batch 中的第 i 个输入样本, y_i 为经分类后的输出类别标签,而 $p(y_i|x_i)$ 为 x_i 被识别为 y_i 类的预测概率。

2. 图像匹配问题:通过分析两张行人图片是否属于同一行人以进行二分类学习。其中常见的损失函数为对比损失(Contrastive Loss),其公式为

$$\mathcal{L}_{con} = (1 - \delta_{ij}) \{\max(0, \rho - d_{ij})\}^2 + \delta_{ij} d_{ij}^2 \quad (2-2)$$

其中 d_{ij} 为两个输入样本 x_i 和 x_j 的嵌入特征之间的欧氏距离; δ_{ij} 为二进制标签指示器,当 x_i 和 x_j 属于同一标签时, $\delta_{ij} = 1$, 否则 $\delta_{ij} = 0$; ρ 为边距参数。

验证损失函数也可以通过交叉熵计算,为二值交叉熵损失,其公式为

$$\mathcal{L}_{veri}(i, j) = -\delta_{ij} \log(p(\delta_{ij}|f_{ij})) - (1 - \delta_{ij}) \log(1 - p(\delta_{ij}|f_{ij})) \quad (2-3)$$

其中 f_i 和 f_j 为两个输入样本 x_i 和 x_j 的嵌入特征,通过计算 $(f_i - f_j)^2$ 可得出差分特征 f_{ij} , 而 $p(\delta_{ij}|f_{ij})$ 为输入样本 x_i 和 x_j 被识别为被识别为 $\delta_{ij} = 0$ 或 1 的概率。

3. 图像检索问题: 网络通过分析多个行人图像, 缩小同一个行人图片的特征距离, 并增大不同行人之间的照片的特征距离。其中三元损失公式如下:

$$\mathcal{L}_{\text{tri}}(i, j, k) = \max(\rho + d_{ij} - d_{ik}, 0) \# (2 - 4)$$

其中 i 为锚点样本, j 为正样本, k 为负样本, 其中锚点样本和正样本为同类的不同样本。 d_{ij} 和 d_{ik} 分别表示锚点样本分别与正样本和负样本的欧式距离, ρ 为边距参数。

2.2 Transformer

谷歌于 2017 年发布 Transformer 网络架构^[39], 最早被用来处理自然语言处理任务。相较于 RNN、LSTM 和 GRU 的顺序处理, Transformer 实现并行化计算, 并且能学习长期依赖关系, 因此在任务中取得良好效果。知名大型语言模型如 GPT-4, LaMDA, LLaMA 和文心一言都基于 Transformer 网络架构。Transformer 最大创新点在于它使用自注意力机制(Self Attention)。

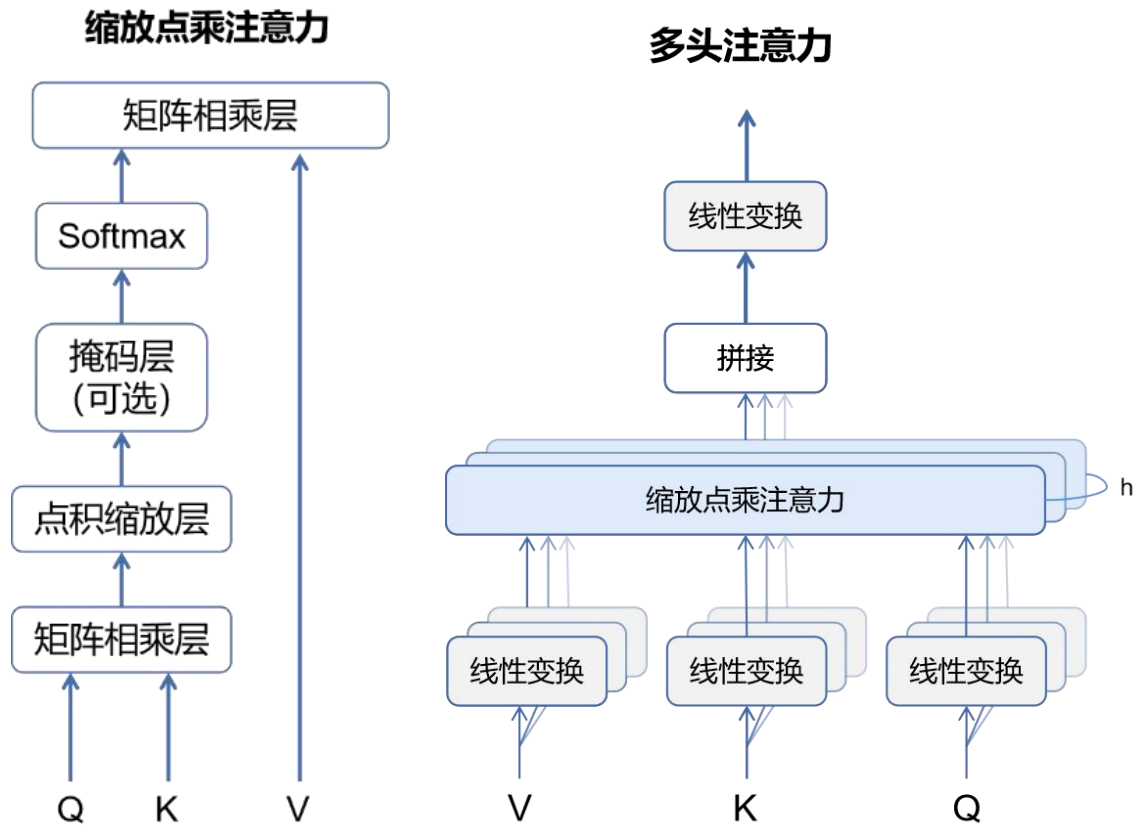
2.2.1 自注意力机制原理

自注意力机制是 Transformer 网络架构中最核心的组成部分。自注意力机制包含矩阵相乘层、点积缩放层、可选的掩码层和交叉熵层(Softmax)。输入向量分别乘上 W^q , W^k 和 W^v 三个不同的权值矩阵, 得到查询向量(Query), 键向量(Key)和值向量(Value)。首先, 查询向量和键向量通过矩阵相乘层计算两者之间的相似性, 并通过点积缩放层调节相似度缩放比例, 以保证参数训练过程不会因为内积值太大而导致难以收敛。交叉熵层将结果归一化后, 值向量与归一化的结果通过矩阵相乘层计算出最终结果。其注意力机制的计算公式为

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \# (2 - 5)$$

若输入向量只与一组权值矩阵相乘, 所得的单一输出难以对实际复杂的应用环境进行准确表征。因此, 网络需要针对不同环境与任务注意不同 token 向量之间的关联性。Transformer 网络架构设置多组可学习权值矩阵来应对复杂的实际情况, 以此形成多头注意力机制(Multi-Head Attention), 其计算公式为

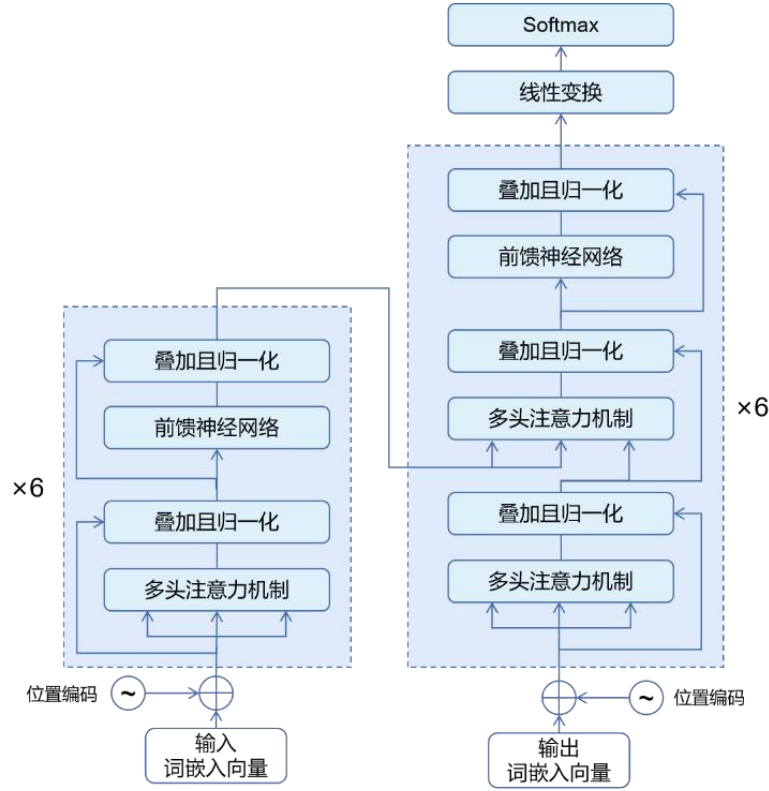
$$\begin{aligned} \text{MultiHead}(Q, K, V) &= \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^0 \# (2 - 6) \\ \text{head}_i &= \text{Attention}(QW^q, KW^k, VW^v) \# (2 - 7) \end{aligned}$$

图 2-1 缩放点乘注意力（左）与多头注意力（右）示意图^[39]

2.2.2 Transformer 网络架构

Transformer 网络架构分为编码器(Encoder)和解码器(Decoder), 都分别由六个相同的编码层和解码层堆叠而成。编码层和解码层基本架构相似, 都由位置编码、前馈神经网络、层归一化和多头注意力机制(Multi-head Attention)组成。由于 Transformer 并没有顺序结构和卷积层, 因此需要额外信息表示信息在序列中的相对和绝对位置。位置编码 (Position Embedding) 维度与输入向量相同, 因此两者可以直接相加。位置编码可以为固定值或由网络自主学习。

为了解决梯度消失的问题, 两者都在每个子网络层(sub-layer)内加入与 ResNet^[40]类似的残差跳跃连接。解码层相比编码层多了一个子网络层, 额外子网络层的多头注意力机制输入为编码器的输出。此外, 编码层使用无掩码层的自注意力机制。由于信息预测时无法得知预测输出后的信息, 因此解码层第一个子网络层使用有掩码器的自注意力机制, 遮挡第 i 个输出后的信息。

图 2-2 Transformer 编码层与解码层架构图^[39]

2.3 视觉 Transformer

受 Transformer 网络架构启发，谷歌在 2020 年发布适用于计算机视觉任务的 ViT 网络^[41]。ViT 相较完整 Transformer 网络架构只使用 Transformer 的编码器。为了尽可能不破坏 Transformer 的网络架构，该文献将图像序列化，把完整输入图像分割成多个大小相等、正方形图像块，形成展平化的二维图像块序列，其中图像解析度为(H, W)，图像块解析度为 (P, P)。图像块数量为

$$N = \frac{HW}{P^2} \#(2 - 8)$$

ViT 中的图像块嵌入加入了类似于 BERT^[42]的可学习[cls]嵌入，以表示图像全局特征。图像块嵌入加入可学习的一维位置编码，以保留位置信息。尽管 ViT 解决了 CNN 的一些缺点，但多头注意力机制的使用让 ViT 有计算与计算开销较大和训练时间较长的缺点。研究人员提出等基于 ViT 的视觉 Transformer，如 Swin Transformer^[43]、DAT^[44]、Deformable DETR^[45]，从不同角度解决上述缺点。

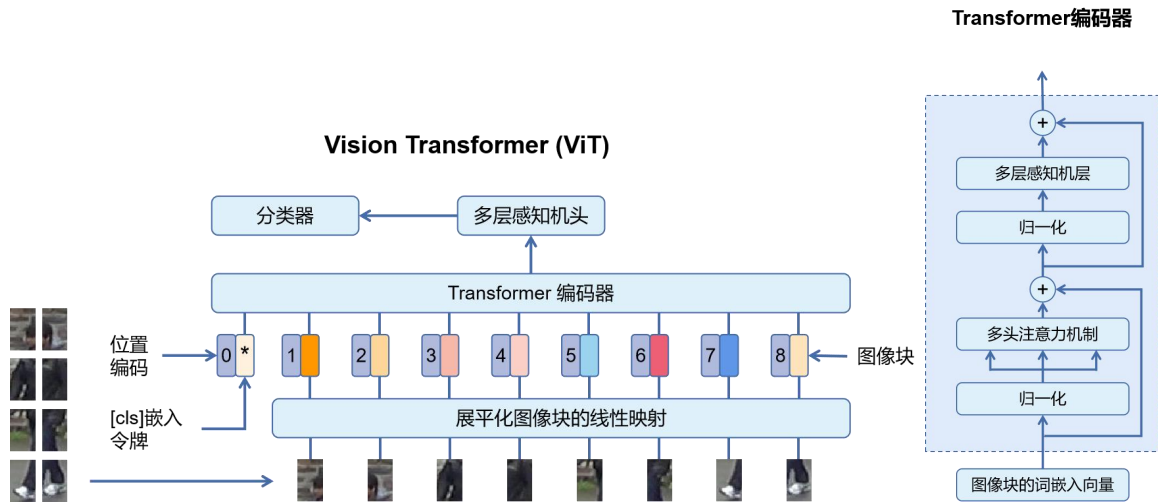


图 2-3 ViT 网络架构示意图（左）Transformer 编码器结构图（右）^[41]

2.4 TransReID 网络

TransReID 是第一个纯视觉 Transformer 行人重识别网络。在这之前，业界普遍使用 CNN 作为行人重识别任务的主干网络。除了主干网络的更动，TransReID 也在输入和输出做出创新。在输入方面，ViT 将完整输入图像分割成多个不互相重叠的图像块，而 TransReID 将完整输入图像分割成多个部分重叠像素的图像块，以保留图像块之间的局部相邻结构。

除了合并输入图像块序列嵌入、网络自主学习的位置编码和[cls]嵌入令牌，TransReID 的输入也包含视角信息编码 (Side Information Embedding)。视角信息嵌入包含非视觉信息，如摄像头 ID 和视点，以帮助视觉 Transformer 学习与摄像头和视点无关的人体特征表示，有效减轻学习特征偏差。行人重识别的图像解析度与图像分类的图像解析度不同，因此网络无法直接使用预训练的位置编码。为了解决这个问题，位置编码将通过双线性 2D 插值，以适应不同解析度的图像。[cls]嵌入令牌则代表全局特征表示。

在输出特征序列处理方面，TransReID 整合特征块重排序模块。此模块通过移位与随机排序的操作打乱特征序列，并将特征序列重组成四个部分。这四个部分都经过最后一层的 ViT 编码层，以学习各个部分的细粒度局部特征。特征块重排序模块利用视觉 Transformer 的全局依赖性，为网络引入额外扰动，提高 TransReID 的鲁棒性、

抗干扰能力和特征辨识能力。

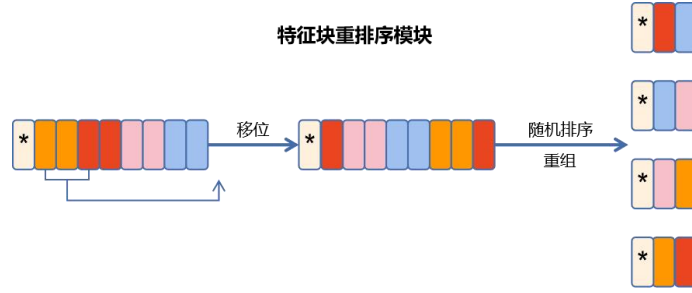


图 2-4 特征块重排序模块示意图^[8]

在损失函数方面，网络结合使用软间隔三元组损失（Triplet Loss with soft-margin）和没有标签平滑的交叉熵损失（ID loss without label smoothing），训练全局特征和局部特征，总损失公式为

$$\mathcal{L} = \mathcal{L}_{ID}(f_g) + \mathcal{L}_T(f_g) + \frac{1}{k} \sum_{j=1}^k \left(\mathcal{L}_{ID}(f_l^j) + \mathcal{L}_T(f_l^j) \right) \#(2-9)$$

其中 f_g 为全局特征， f_l^j 为第 j 个局部特征， k 为局部特征数量。

由于 TransReID 使用 ViT 作为主干网络，并且不依靠复杂的辅助信息处理网络，因此本文选定 TransReID 作为基线网络，在 TransReID 的基础上新增各个模块，并与本文提出的网络进行性能比较。

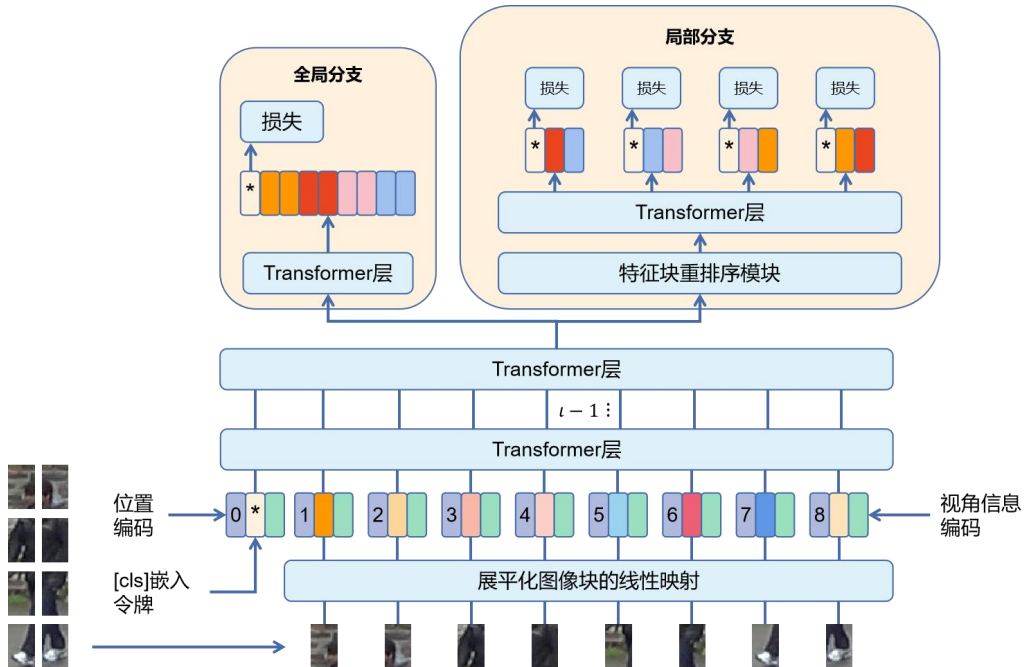


图 2-5 TransReID 网络架构示意图^[8]

2.4 本章小结

本章初步介绍行人重识别数据集与行人重识别技术大致实现方向。之后，本章阐述以自注意力机制为核心的 Transformer 网络架构和其底层逻辑。研究人员受到 Transformer 的启发后以此发布适用于计算机视觉领域的 ViT 网络，其主要特点为图像序列化，并且相比 CNN 有诸多好处，但 ViT 也有自己的缺点。最后，本章阐述基于视觉 Transformer 的行人重识别网络 TransReID，并将 TransReID 作为基线网络，作为本文网络的对照。TransReID 的创新点包括部分重叠图像块的使用，视觉信息编码的加入和特征块重排序模块的应用，以提升网络性能。

第三章 基于数据增强与特征融合的行人重识别网络

3.1 网络架构

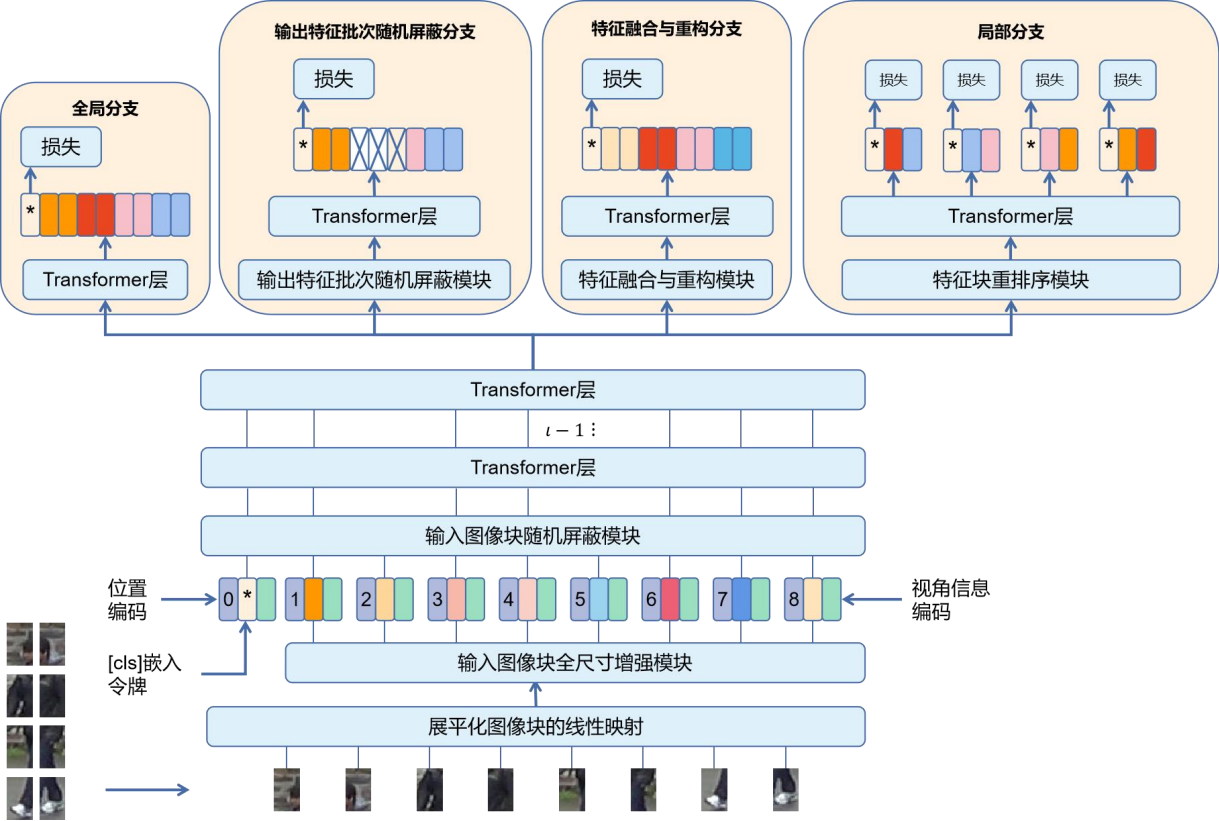


图 3-1 网络架构示意图

图 3-1 为本文提出的行人重识别网络架构。输入图像被分割成多个图像块后进行展平，形成输入图像序列。经过输入图像块全尺寸增强模块后，输入图像序列加入各个辅助信息，包括位置编码，[cls]嵌入令牌和视角信息编码。输入图像块随机屏蔽模块将随机屏蔽完整输入图像序列，并送入 ViT 主干网络。ViT 主干网络共有 12 层 Transformer 编码器。

经过 11 层 Transformer 编码层后，输出特征序列将被送入四个分支，分别是全局特征分支、输出特征批次随机屏蔽分支、特征融合与重构分支和局部特征分支。全局特征分支的特征并不会经过其他模块，并使用 ViT 主干网络中提取的完整输出特征序

列进行分类并计算损失。其余分支的特征将分别经过输出特征批次随机屏蔽模块、特征融合与重构模块和特征块重排序模块后,使用经处理的输出特征序列进行分类并计算损失。

3.1.1 行人重识别任务主干网络的选择

ViT 网络架构的出现为研究人员提供新的思路处理计算机视觉任务。CNN 受限的感受野导致卷积神经网络只能专注于提取范围较小判别区域。相比之下视觉 Transformer 并没有使用卷积层、池化层、下采样层等,因此输入图像可以保留完整信息,以此提高网络的特征辨识能力,识别两个特征相似但身份不同的行人。此外,CNN 无法解决长距离依赖问题。即使加入注意力模块,但感受野受限导致注意力模块不适用于连续且大的区域,因此也很难提取判别性较强的特征。综上所述,本文使用基于视觉 Transformer 的行人重识别网络 TransReID 作为基线网络。

3.1.2 网络新增模块

本文提出的行人重识别网络在 TransReID 网络的基础上新增四个模块,其中三个模块与数据增强有关,分别是输入图像块随机屏蔽模块、输入图像块全尺寸增强模块和输出特征批次随机屏蔽模块。另一个模块和特征融合有关,为特征融合与重构模块。

3.2 输入图像块全尺寸增强模块

行人图像本身就具有复杂的背景信息和干扰,如遮挡物、其他行人和行人所处的采集场景。若不处理这些背景信息,网络将学习这些不属于行人的特征,从而降低网络性能。其中最为直接的方法为假设行人重识别数据集中所有行人图像的正中央为行人,而图像边缘区域为背景和噪音。这样的假设并无法适应现实环境的行人图像采集,因行人并不会只出现图像正中央。网络动态调整需要着重关注的区域,并减弱其他背景与噪音对特征提取的影响是具有研究意义的。

3.2.1 具体实现方法

参考此文献^[9],输入图像块全尺寸增强模块作用在无[cls]嵌入令牌、位置编码和视角信息嵌入的输入图像块序列,因为这些辅助信息并不带有噪音,且这些信息不存在重要性排序,因此无需对这些信息进行增强或减弱。此模块部署在输入图像块序列

随机屏蔽模块之前。模块生成与输入图像块序列大小一致的可学习张量，其中大小为 $P \times D$ ， P 为输入图像块数量， D 为每个图像块的维度。此张量所有元素的初始化值为 1。输入图像块序列与可学习张量作哈达玛积，作为视觉 Transformer 新的输入图像块序列。

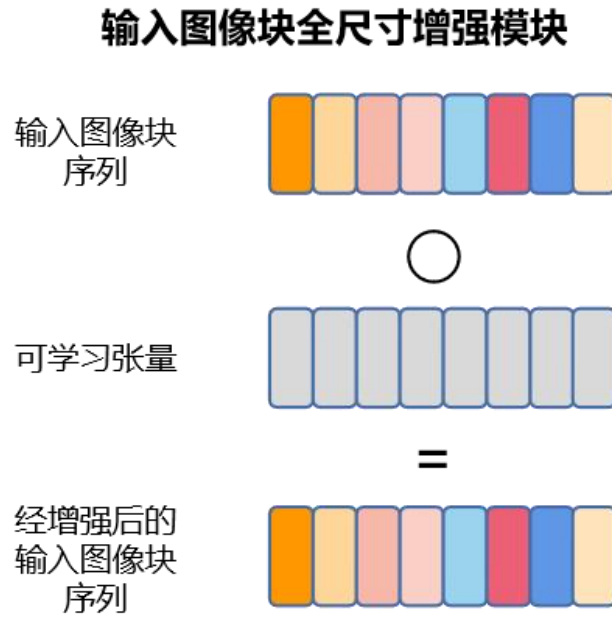


图 3-2 输入图像块全尺寸增强模块示意图

可学习张量对信息含量较少或带有噪音的图像块乘上小于 1 的值，降低这些图像块在网络中的影响，相当于降低图像噪音；并对信息含量较大的图像块乘上大于 1 的值，增强这些图像块在网络中的影响。

此模块能在小幅增加训练时间、计算与显存开销的前提下增强输入图像块序列的特征表示能力，提升识别准确率和降低噪音影响。此模块能轻松部署在不同基于视觉 Transformer 的行人重识别网络中。

3.2.2 实验数据

为了验证此模块的效果，表 3-1 和表 3-2 分别为 DukeMTMC-reID 数据集和 Occluded Duke 数据集的实际实验数据。步长(stride size)为 16，图像解析度为 256×128 。根据下表数据，无论输入图像块随机屏蔽模块是否开启，此模块可以有效提高网络性能。

此外，此模块对遮挡行人重识别数据集效果更好。根据表 3-2 的数据，当被屏蔽

的图像块越多，则性能提升越明显。相较于 $rate_{keep} = 1.0$ 时性能提升幅度为 $+0.1\% mAP$ ，当 $rate_{keep} = 0.5$ 时性能提升幅度为 $+0.8\% mAP$ ，随着输入图像块数量的减少和位置分布上的分散，可学习张量的变化会更加明显，进而更容易突出含有重要信息的图像块。因此，此模块与输入图像随机屏蔽模块可以共同提升网络性能与效率。

表 3-1 DukeMTMC-reID 数据集在输入图像块随机屏蔽模块的性能指标

$rate_{keep}$	输入图像块 序列全尺寸 增强模块	mAP (%)	R-1 (%)	R-5 (%)	R-10 (%)
1.0	关闭	80.3	89.8	95.6	96.8
	开启	80.5	89.5	96.0	97.1
0.7	关闭	79.4	88.6	95.0	96.6
	开启	79.5	89.1	94.9	96.3

表 3-2 Occluded Duke 数据集在输入图像块随机屏蔽模块的性能指标

$rate_{keep}$	输入图像块 序列全尺寸 增强模块	mAP (%)	R-1 (%)	R-5 (%)	R-10 (%)
1.0	关闭	56.0	64.0	79.0	84.5
	开启	56.1	64.6	79.5	84.2
0.7	关闭	52.4	58.7	73.9	80.1
	开启	52.7	59.5	74.7	80.7
0.5	关闭	48.1	53.3	69.4	75.7
	开启	48.9	53.7	70.7	77.1

此外，本文针对此模块的部署位置进行实验验证，决定此模块是否需要增强位置编码和视角信息嵌入。表 3-3 和表 3-4 分别为 DukeMTMC-reID 数据集和 Occluded Duke 数据集的实际实验数据。步长为 16，图像解析度为 256×128 。根据下表数据，辅助信息的增强反而导致网络性能下降。各个图像块的辅助信息都是同等重要，并且

这个辅助信息都不是噪音,因此对这些辅助信息进行增强或减弱这些辅助信息将影响这些信息对网络性能的积极影响。综上所述,输入图像块全尺寸增强模块无需增强辅助信息。

表 3-3 DukeMTMC-reID 数据集在辅助信息增强下的性能指标

$rate_{keep}$	辅助信息增强	mAP (%)	R-1 (%)	R-5 (%)	R-10 (%)
1.0	有	79.3	89.3	95.0	97.0
	无	80.5	89.5	96.0	97.1

表 3-4 Occluded Duke 数据集在辅助信息增强下的性能指标

$rate_{keep}$	辅助信息增强	mAP (%)	R-1 (%)	R-5 (%)	R-10 (%)
1.0	有	53.4	61.5	77.5	82.6
	无	56.1	64.6	79.5	84.2

3.3 输入图像块随机屏蔽模块

相较于基于 CNN 的行人重识别算法,基于视觉 Transformer 的行人重识别算法蕴含巨大潜能,但视觉 Transformer 的计算与显存开销较高。研究人员提出通过修改视觉 Transformer 的架构或复杂的训练流程的方式以提高网络效率,但这些修改都限制了视觉 Transformer 应用场景。因此,研究一种通用于视觉 Transformer 的效率提高模块是有意义的。通过随机屏蔽一定数量输入图像块,视觉 Transformer 的输入将减少,从而提升网络效率。

3.3.1 具体实现方法

参考此文献^[46],输入图像块随机屏蔽模块处理 TransReID 完整输入图像块序列,包括位置编码、[cls]嵌入令牌和视角信息嵌入。模块先分别构建与[cls]嵌入令牌和输入图像块序列维度一致的掩膜,而[cls]嵌入令牌默认被保留。此外,输入图像块序列保留比例 $rate_{keep}$ 和输入图像块序列长度之积 $patch_{keep}$ 将是需要保留的输入图像块数量。此模块将生成一个大小与输入图像块序列一致的掩膜。模块对输入图像块序列掩膜随机生成任意数值后,输入图像块序列掩膜将按照顺序重新排列,并保留前

$patch_{keep}$ 个输入图像块序列。此模块在网络训练期间正常运作，而在网络测试期间关闭模块。

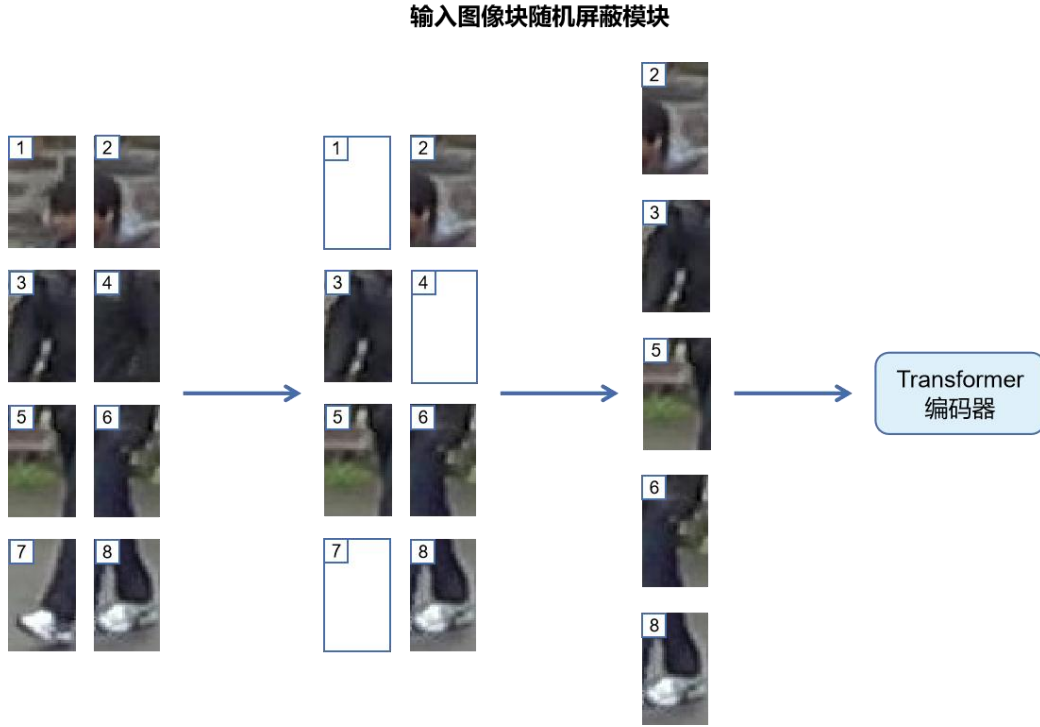


图 3-3 输入图像块随机屏蔽模块示意图

3.3.2 现有方法对比

此模块与普遍使用的数据增强方法不同。两者虽然都可以增加数据的多样性，正则化数据，减少过拟合的风险，但两者并不作用在网络的同个部分。首先，数据增强作用在完整图像，之后图像才会被切割成多个输入图像块并作为 ViT 网络的输入，图像块数量并没有减少，因此数据增强虽然可以一定程度上增强网络性能，但依旧无法提高网络效率、减少训练时间。

输入图像块随机屏蔽模块作用在输入图像块序列，通过随机屏蔽图像块让 ViT 主干网络处理更少图像块，从而加快训练速度。由于部分图像块所蕴含的信息并不多，因此 ViT 主干网络随机忽略这些图像块并不会造成网络性能的剧烈下降，反而可以加快训练速度。尽管如此，主干网络无法在尚未开始进行特征提取的前提下分析每个图像块的信息含量。随机屏蔽一部分图像块的做法无需提前对每个图像块的信息含量进行假设，更为简单。

3.3.3 实验数据

为了观察 $rate_{keep}$ 对性能、训练时间和计算开销影响, 表 3-5 至 3-7 分别是 DukeMTMC-reID 数据集、Market-1501 数据集和 Occluded Duke 数据集的实际实验数据。此实验并没有运用其他模块, 步长为 16, 图像解析度为 256×128 。

表 3-5 DukeMTMC-reID 数据集在不同 $rate_{keep}$ 下的性能指标与训练开销

$rate_{keep}$	训练时间	mAP (%)	R-1 (%)	R-5 (%)	R-10 (%)	训练时间变化 (%)	性能变化 (%)	训练所用显存 (MiB)	训练显存变化 (%)
1.0	3:41:52	80.3	89.8	95.6	96.8	—	—	5171	—
0.8	3:06:07	80.1	89.6	95.2	96.8	-21.75	-0.2	4439	-14.16
0.7	2:48:00	79.4	88.6	95.0	96.6	-29.37	-0.9	4181	-19.15
0.5	2:16:52	77.9	87.1	94.1	95.8	-42.46	-2.4	3745	-27.58

表 3-6 Market-1501 数据集在不同 $rate_{keep}$ 下的性能指标与训练开销

$rate_{keep}$	训练时间	mAP (%)	R-1 (%)	R-5 (%)	R-10 (%)	训练时间变化 (%)	性能变化 (%)	训练所用显存 (MiB)	训练显存变化 (%)
1.0	2:59:53	87.6	94.4	98.2	99.0	—	—	5095	—
0.8	2:30:25	87.3	94.3	98.3	99.0	-16.38	-0.3	4525	-11.19
0.7	2:14:23	87.0	94.6	98.3	99.0	-25.29	-0.6	4217	-17.23
0.5	1:49:21	85.1	93.7	97.7	98.7	-39.21	-2.5	3773	-25.95

表 3-7 Occluded Duke 数据集在不同 $rate_{keep}$ 下的性能指标与训练开销

$rate_{keep}$	训练时间	mAP (%)	R-1 (%)	R-5 (%)	R-10 (%)	训练时间变化 (%)	性能变化 (%)	训练所用显存 (MiB)	训练显存变化 (%)
1.0	3:36:07	56.0	64.0	79.0	84.5	—	—	5131	—
0.8	2:52:19	54.5	61.2	77.4	82.3	-20.27	-1.5	4439	-13.49
0.7	2:39:27	52.4	58.7	73.9	80.1	-26.22	-3.6	4245	-17.27
0.5	2:13:14	48.1	53.3	69.4	75.7	-38.35	-7.9	3799	-25.96

根据实验数据, 随着 $rate_{keep}$ 的下降, 性能变化的下降相较训练时间变化和计算

开销的下降来得快。此外,表 3-7 的数据表示此模块能在遮挡行人重识别数据集中发挥作用。在 $rate_{keep}$ 参数的选择上,尽管使用 $rate_{keep} = 0.8$ 的网络相比使用 $rate_{keep} = 0.7$ 的网络有更高的效率,但此模块的另一大目的是数据集正则化,通过减少输入图像块的数量模拟遮挡行人重识别数据集,以缓解数据数量不足的困境。 $rate_{keep} = 0.8$ 并不足以完成上述目的,让网络学习更多具有辨识度的特征。综上所述,本文将 $rate_{keep}$ 设定在 0.7,并在后续实验保持该参数的不变。

其中一个非常直接能有效提升网络性能的方法是让网络处理更多图像块。目前有两种方式能从一张完整图像获取更多图像块,一是通过减小步长大小,二是提高图像解析度。步长大小越小,各个图像块大小也会越小,则图像块数量越多。由于每个图像块大小和形状固定,因此即使步长大小固定,当图像解析度越高,图像块数量也会越多。对较为背景复杂的图像进行更为细致的分割可以有助网络学习更具辨识度的特征,从而提升网络性能。

虽然更多图像块确实可以提高网络性能,但相关调整并无法平衡性能和计算与显存开销,因此虽然原 TransReID 论文针对上述方法进行实验,但最终并没有采用此方案。此模块可减少网络训练时长、计算与显存开销,而被节省的时间和计算与显存开销可以让网络处理更多与完整图像占比较小的图像块。为了验证上述说法,表 3-8 至 3-10 分别是三个不同数据集在不同步长下的实际实验数据,图像解析度固定在 256×128 。

根据实验数据,减少步长大小的确可以在消耗更多训练时间与计算开销的前提下提高性能。输入图像块随机屏蔽模块和调整步长大小的结合成功让网络在更短的训练时间和计算开销的前提下达成与原网络相同的性能,或在大致相同的训练时间和计算开销的前提下达成更高的性能。

表 3-8 DukeMTMC-reID 数据集在不同步长下的性能指标与训练开销

$rate_{keep}$	步长	训练时间	mAP (%)	R-1 (%)	R-5 (%)	R-10 (%)	训练时间变化 (%)	性能变化 (%)	训练所用显存 (MiB)	训练显存变化 (%)
1.0	16	3:41:52	80.3	89.8	95.6	96.8	-	-	5171	-
0.7	16	2:48:00	79.4	88.6	95.0	96.6	-29.37	-0.9	4181	-19.15
0.7	14	3:22:10	80.3	89.6	95.2	96.5	-15.00	+0.0	4709	-8.93
0.7	12	4:04:11	80.8	89.5	95.3	96.9	+2.66	+0.5	5589	+8.08

表 3-9 Market-1501 数据集在不同步长下的性能指标与训练开销

$rate_{keep}$	步长	训练时间	mAP (%)	R-1 (%)	R-5 (%)	R-10 (%)	训练时间变化 (%)	性能变化 (%)	训练所用显存 (MiB)	训练显存变化 (%)
1.0	16	2:59:53	87.6	94.4	98.2	99.0	-	-	5095	-
0.7	16	2:14:23	87.0	94.6	98.3	99.0	-25.29	-0.6	4217	-17.23
0.7	14	2:38:24	87.4	94.3	98.2	99.1	-11.94	-0.2	4925	-3.34
0.7	12	3:14:38	88.2	94.7	98.2	99.0	+8.20	+0.6	5685	+11.58

表 3-10 Occluded Duke 数据集在不同步长下的性能指标与训练开销

$rate_{keep}$	步长	训练时间	mAP (%)	R-1 (%)	R-5 (%)	R-10 (%)	训练时间变化 (%)	性能变化 (%)	训练所用显存 (MiB)	训练显存变化 (%)
1.0	16	3:36:07	56.0	64.0	79.0	84.5	-	-	5131	-
0.7	16	2:39:27	52.4	58.7	73.9	80.1	-26.22	-3.6	4245	-18.88
0.7	14	3:08:59	54.7	61.4	77.9	83.2	-12.55	-1.3	4763	-8.98
0.7	12	3:47:11	56.4	63.3	79.0	84.6	+5.12	+0.4	5627	+7.53

除了通过调整步长大小，增大输入图像解析度也可以提升网络性能。根据原文献的描述^[46]，相比直接使用原始大小的图像，使用 16 倍大小的图像并只保留 5% 的图像块可以提升性能并节省计算开销和显存占用。为了观察图像解析度大小对网络性能、训练时间和显存占用的影响，表 3-11 至 3-13 分别是三个不同数据集在不同图像解析度下的实际实验数据。

根据下表数据，图像解析度的提高确实可以在消耗更多训练时间与计算开销的前提下提高性能。尽管如此，随着图像解析度和步长的提高，网络性能的上升幅度有所下降，因此网络效率反而有所降低。单纯使用更高解析度照片或使用更小步长大小虽然可以进一步提升性能，但网络效率仍然不及原 TransReID 网络。综上所述，输入图像块随机屏蔽模块并不足以同时满足提升网络性能与提高网络效率两项目标，因此网络需要使用其他模块以增强网络性能。根据下表数据，为了在性能、训练时间与计算开销达成平衡，最终网络参数固定在 $rate_{keep} = 0.7$ ，步长大小 14，图像解析度 384×128 。

表 3-11 DukeMTMC-reID 数据集在不同步长与图像解析度下的性能指标与训练开销

$rate_{keep}$	步长	图像解析度	训练时间	mAP (%)	R-1 (%)	R-5 (%)	R-10 (%)	训练时间变化 (%)	性能变化 (%)	训练所用显存 (MiB)	训练显存变化 (%)
1.0	16	256×128	3:41:52	80.3	89.8	95.6	96.8	—	—	5171	—
0.7	16	384×128	3:50:09	80.1	89.1	95.2	96.9	-3.82%	-0.2	5517	+6.69
0.7	14	384×128	4:37:58	81.0	89.9	95.4	97.0	+16.16	+0.7	6475	+16.31
0.7	12	384×128	5:38:09	81.3	89.9	95.6	96.9	+41.31	+1.0	7693	+38.19

表 3-12 Market-1501 数据集在不同步长与图像解析度下的性能指标与训练开销

$rate_{keep}$	步长	图像解析度	训练时间	mAP (%)	R-1 (%)	R-5 (%)	R-10 (%)	训练时间变化 (%)	性能变化 (%)	训练所用显存 (MiB)	训练显存变化 (%)
1.0	16	256×128	2:59:53	87.6	94.4	98.2	99.0	—	—	5095	—
0.7	16	384×128	3:04:56	86.8	93.9	98.1	99.0	+1.7	-0.8	5453	+7.03
0.7	14	384×128	3:44:36	87.9	94.4	98.2	99.0	+23.52	+0.3	6465	+26.89
0.7	12	384×128	4:45:45	88.1	94.6	98.2	99.0	+57.15	+0.5	7733	+51.78

表 3-13 Occluded Duke 数据集在不同不同步长与图像解析度下的性能指标与训练开销

$rate_{keep}$	步长	图像解析度	训练时间	mAP (%)	R-1 (%)	R-5 (%)	R-10 (%)	训练时间变化 (%)	性能变化 (%)	训练所用显存 (MiB)	训练显存变化 (%)
1.0	16	256×128	3:36:07	56.0	64.0	79.0	84.5	—	—	5131	—
0.7	16	384×128	3:39:22	54.5	60.9	77.1	82.5	+1.50	-1.5	5337	+4.01
0.7	14	384×128	4:25:19	57.5	64.6	80.4	85.1	+22.77	+1.5	6353	+23.82
0.7	12	384×128	5:27:57	57.9	65.7	79.9	85.2	+51.75	+1.9	7547	+47.09

3.4 输出特征批次方块随机屏蔽模块

基于 CNN 的行人重识别网络着重于主要身体部位，而忽视其他次要身体部位。

为了让网络也关注这些次要身体部位的特征,研究人员提出多个基于局部身体部位和姿势的网络,分割不同身体部位并分别进行特征学习,但这些方法都需要额外的身体姿势信息。因此研究一个通用,不依靠其他额外信息且可以完成上述要求的模块是研究目标之一。

参考此文献^[47],输出特征批次方块随机屏蔽模块通过对同一批次的特征图进行相同位置的特征屏蔽,相当于在同一批次的特征图中删除相同语义的身体部位。对输出特征序列的操作能增强局部区域的关注特征学习,增强网络的泛化能力与性能。特征图掩膜大小为

$$size_{mask} = h \times ratio_h \times w \#(3 - 1)$$

其中 h 为完整特征图高度, $ratio_h$ 为高度比, w 为完整特征图宽度。尽管模块在不增加网络规模的前提下提升性能,但该模块直接处理完整图像,而不是输入图像块序列,因此该模块仅适用于 CNN,无法直接套用在 ViT 里。参考原论文的思路,本文提出适用于 ViT 的输出特征批次方块随机屏蔽模块。

3.4.1 具体实现方法

此模块以 ViT 主干网络第 11 层的输出特征图作为输入,让 ViT 主干网络的最后一层从被随机屏蔽的输出特征图提取更具辨识度的特征。输入图像块总数量为完整图像纵向图像块数量 $patch_x$ 和完整图像横向图像块数量 $patch_y$ 之积,而输入图像块总数量与输出特征块数量相等。 $patch_y$ 在完整图像中可被视为完整图像高度,而 $patch_x$ 可被视为完整图像宽度。

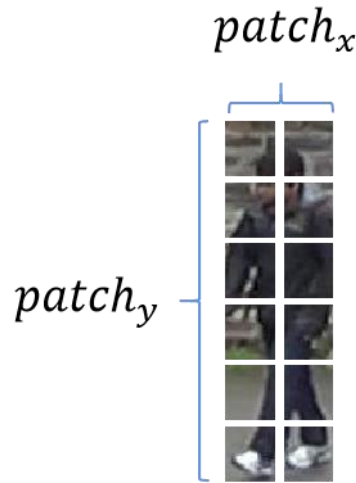


图 3-4 输入图像块总数量示意图

特征图掩膜的纵向图像块数量为

$$mask_y = patch_y \times ratio_h \#(3-2)$$

特征图掩膜的横向图像块数量 $mask_x$ 与 $patch_x$ 相等，特征图掩膜大小为

$$size_{mask} = mask_x \times mask_y \#(3-3)$$

模块计算得出特征图掩膜纵向图像块数量后将随机在 $[0, (patch_y - mask_y - 1)]$ 中选择特征图掩膜纵向起点，称为 $mask_{start}$ 。特征图掩膜在完整特征图中的起点特征图像块和终点特征图像块为 $[(mask_{start} \times patch_x), (mask_{start} + mask_y) \times patch_x]$ 。 $[cls]$ 嵌入令牌默认不会被屏蔽，以持续传递此前 ViT 网络学习的全局特征。 $[cls]$ 嵌入令牌掩膜与特征图掩膜合并，并对完整特征图进行哈达玛积 (Hadamard Product)，获得经掩膜的输出特征序列。主要模块框架如下图所示。

输出特征批次方块随机屏蔽模块

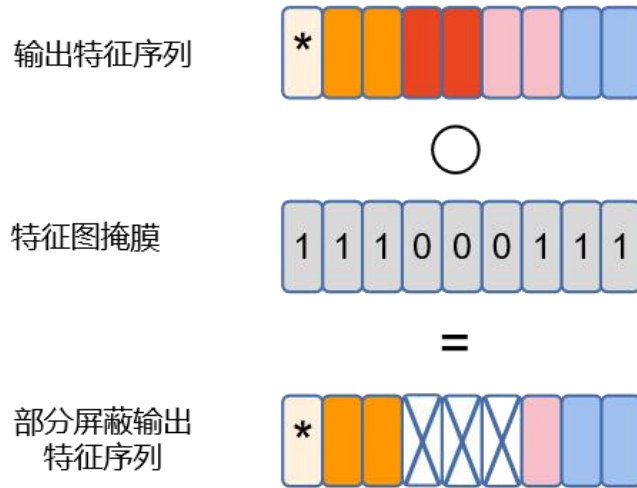


图 3-5 输出特征批次方块随机屏蔽模块示意图

3.4.2 现有方法对比

本文提出适用于 ViT 主干网络的输出特征批次方块随机屏蔽模块相较于原论文在网络架构层面仍然有差别。首先，原模块作用在经过 ResNet 第四阶段完成特征提取的完整特征图。ResNet 第四阶段并不进行降采样操作，而是将特征图进行瓶颈处理 (bottleneck)，再对特征图进行随机屏蔽。特征图经过全局最大池化，让网络能在被屏蔽具有辨识度特征的特征图中凸显此前被忽略的特征，最后才分别经过相较之下，ViT 网络架构决定本文提出的模块并没有使用全局最大池化。此外，瓶颈处理作用在

经过随机屏蔽后的输出特征序列。

此外，此模块与随机失活 (Dropout) 不同。虽然两者都针对输出特征图进行随机屏蔽，也都可以对数据或特征正则化以防止网络过拟合，但随机失活模块对每张照片的随机部位进行像素级屏蔽，并无法遮挡相同语义身体部位，进而无法增强局部区域的关注特征学习。对同一批次的图像进行不同位置的屏蔽只能达成数据集正则化，并无法让网络聚焦于学习不被遮挡的身体部位的特征，因此本文提出的输出特征批次方块随机屏蔽模块可以同时完成数据集正则化和增强局部区域的特征学习。

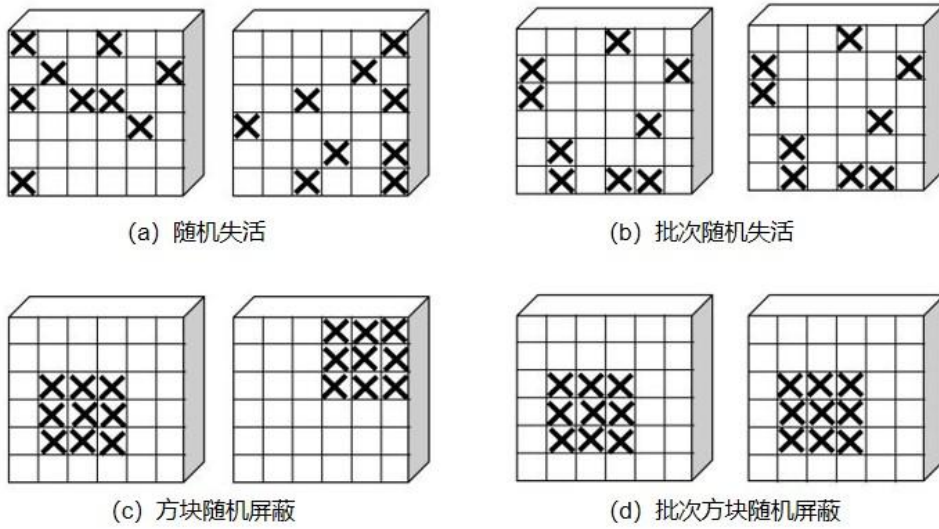


图 3-6 输出特征批次方块随机屏蔽模块与其他随机屏蔽模块对比示意图

3.4.3 实验数据

根据原文献^[47]，输出特征批次方块随机屏蔽模块的 $ratio_h$ 应选择 0.3，以达到网络最佳性能。此外，输出特征批次方块随机屏蔽模块应部署在主干网络最后一层前，让主干网络的最后一层通过被屏蔽的特征图进行特征学习。

本文先验证原 $ratio_h$ 的合理性，表 3-14 为 DukeMTMC-reID 数据集在不同 $ratio_h$ 下的实际实验数据，其中此实验没有运用其他模块，步长大小为 16，图像解析度为 256×128 。此模块被部署在 ViT 主干网络第 11 层之后。表 3-15 则是 Market-1501 数据集在不同 $ratio_h$ 下的实际实验数据，其中 $rate_{keep}$ 为 0.7，其他网络设定与上述一致。

根据下表数据，不同数据集应当选用不同的 $ratio_h$ 值，以最大化提升网络性能，这与数据集特性有一定关系。综上所述，DukeMTMC-reID 数据集选用的 $ratio_h$ 为 0.3，Market-1501 数据集选用的 $ratio_h$ 为 0.2。

表 3-14 DukeMTMC-reID 数据集在不同 $ratio_h$ 下的性能指标

$ratio_h$	mAP (%)	R-1 (%)	R-5 (%)	R-10 (%)
0.2	80.5	89.5	95.4	96.8
0.3	80.6	89.8	95.5	97.1
0.4	80.4	89.6	95.2	96.8

表 3-15 Market-1501 数据集在不同 $ratio_h$ 下的性能指标

$ratio_h$	mAP (%)	R-1 (%)	R-5 (%)	R-10 (%)
0.2	87.0	94.6	98.2	99.0
0.3	86.6	94.2	98.1	99.1
0.4	86.9	94.7	98.3	99.1

此外,本文也验证此模块对遮挡行人重识别数据集的效果。下表为 Occluded Duke 数据集在不同 $ratio_h$ 下的实际实验数据。

表 3-16 Occluded Duke 数据集在不同 $ratio_h$ 下的性能指标

$ratio_h$	mAP (%)	R-1 (%)	R-5 (%)	R-10 (%)
0.1	56.7	65.3	79.5	84.6
0.2	55.3	62.5	78.5	83.0
0.3	55.2	62.1	78.1	82.6

当此模块使用相同大小的 $ratio_h$ 时,网络性能反而下降。由于遮挡行人重识别数据集已经部分遮挡行人特征,因此过大的 $ratio_h$ 将屏蔽更多与遮挡物无关的行人特征,降低网络的特征提取能力。

尽管如此,较小的 $ratio_h$ 值能有效提升网络性能,相较基线网络提升 0.7% mAP。较小的 $ratio_h$ 值在不屏蔽过多行人特征的前提下去除图像噪声,提升网络性能。根据实验数据, Occluded Duke 数据集选用的 $ratio_h$ 为 0.1。综上所述,此模块不仅能作为行人重识别数据集的数据增强模块,也可以有效提升网络在遮挡行人重识别数据集的性能。

上述实验确定 $ratio_h$ 后,接下来需要确定输出特征批次方块随机屏蔽模块在整个网络的部署位置。为了观察不同部署位置对目前网络性能的影响,下表为 DukeMTMC-reID 数据集的实际实验数据,其中模块会被部署在 ViT 主干网络的不同编码层之后。根据下表数据,当模块被部署在 ViT 主干网络的第 11 个编码层后,

则网络性能最好。

表 3-17 DukeMTMC-reID 数据集在模块部署在不同 ViT 层数下的性能指标

ViT 编码层	mAP (%)	R-1 (%)	R-5 (%)	R-10 (%)
11	80.6	89.8	95.5	97.1
12	80.4	89.1	95.3	96.8

3.5 特征融合与重构模块

ViT 能有效提取全局特征向量，但现实环境中的行人非常容易受到其他行人和遮挡物的影响。若网络只关注全局特征向量，网络鲁棒性和泛化能力会因为遮挡物而相对减弱，因此网络也必须提取局部特征，增强网络性能。TransReID 中的特征块重排序模块是其中一个方式，但目前局部特征提取与融合并没有考虑到特征序列内部之间的空间相关性。空间相关性的研究让网路能从中提取更多信息，因此相关研究是必要的。

完整图像存在部分范围蕴含大量具有辨识度特征与信息，同样也有部分范围蕴含相对少量特征与信息。通过计算各个局部特征块和全局特征之间的余弦相似度(cosine similarity)，较为重要、有更多行人特征与信息特征大多集中在输出特征序列的中段位置，而位于输出特征序列前段和后段位置的特征块则相对不重要。此外，在输出特征序列前端和后段的特征块普遍上都有背景和其他物件的杂讯，这些信息会一定程度上干扰网络。

尽管如此，这不意味着这些位置的特征块应该被丢弃，因此这些位置的特征块也存在一些有关行人的辅助信息，如头部、脚部和与其相关的物体，帮助网络从中区分不同的行人。由于这些特征块之间不同的重要性，因此通过特征融合与重构模块加强重要特征块的重要性，网络可以从中提取更多信息，降低噪音的影响。

3.5.1 具体实现方法

参考此文献^[9]，输出特征序列分为四等分的分支输出特征序列，其中第一个和第四个序列分别称作头序列和尾序列。 $[cls]$ 嵌入令牌从完整输出特征序列分离，不受模块影响。头尾序列乘上小于 1 的特征融合比例 f_{scale} ，将降低头尾特征序列对输出的影响。头尾序列分别与第二序列和第三序列相加，形成新的头尾序列。最后新的头尾

特征序列与两个主要特征序列和[cls]嵌入令牌进行合并，形成新的输出特征序列。

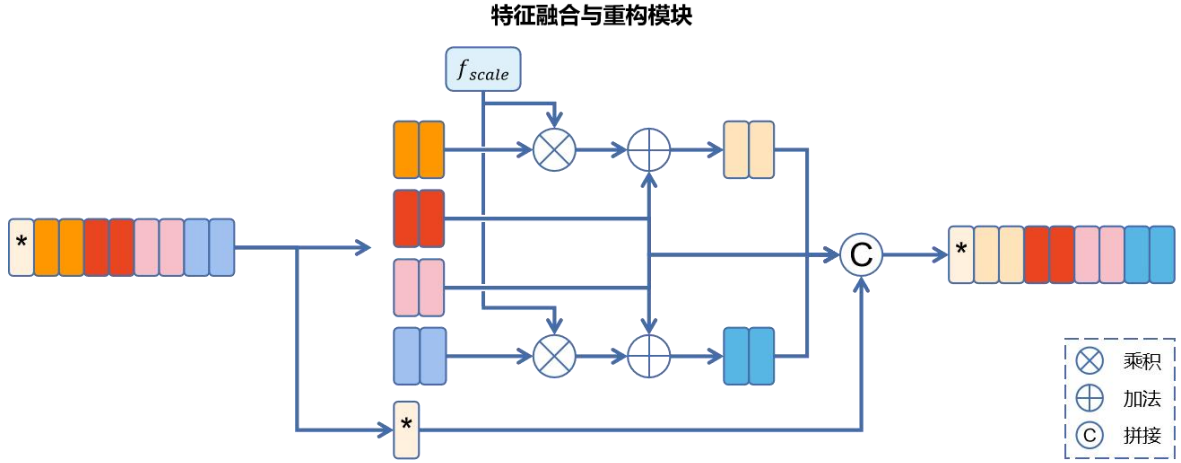


图 3-7 特征融合与重构模块示意图

3.5.2 实验数据

由于融合方案的选择对网络性能影响较大，因此需经过实验并根据性能差异方可决定融合方案。本文尝试三个不同的特征融合方案，以增强重要特征块在网络的影响力。三个方案都先将头尾序列的特征乘上小于 1 的 f_{scale} ，以降低头尾序列在网路的重要性。

- 方案 1 将新的头尾序列分别与第二序列和第三序列相加。
- 方案 2 在方案 1 的基础之上分别对头尾序列进行平均计算。
- 作为对照，方案 3 不做任何后续计算。

为了验证这些融合方案的有效性，表 3-18 为 DukeMTMC-reID 数据集的实际实验数据。此实验并没有运用其他模块，步长为 16，图像解析度为 256×128 。根据下表实验数据，方案 1 整体的性能指标较其他方案来得好。方案 1 从两个方向增强重要特征块的影响力。一方面降低头尾序列的重要性，而另一方面与第二序列和第三序列相加。其他两个方案均减弱重要特征块网络影响力的增强幅度，因此性能提升不如方案 1。

挑选特征融合方案与 f_{scale} 的标准主要为 mAP 和 R-1。由于 mAP 更加全面衡量重识别算法性能，因此本文优先考虑 mAP 性能指标。若多个方案的 mAP 值相同，则会相互比较 R-1，R-5 和 R-10 也以此类推。根据下表数据，基于较高的 mAP，因此 $f_{scale} = 0.5$ 的方案 1 为最终特征融合方案。

表 3-18 DukeMTMC-reID 数据集在不同特征融合方案下的性能指标

特征融合方案	f_{scale}	mAP (%)	R-1 (%)	R-5 (%)	R-10 (%)
方案 1	0.3	80.4	89.7	95.5	97.0
	0.5	80.6	89.5	95.7	97.1
	0.7	80.2	89.4	95.3	97.0
	1.0	80.3	89.7	95.4	96.9
方案 2	0.5	80.2	89.7	95.5	96.9
	0.7	80.3	89.3	96.1	97.2
	1.0	80.1	89.0	95.2	96.9
方案 3	0.3	80.4	89.7	95.4	96.8
	0.5	80.1	89.9	95.4	96.9
	0.7	80.2	89.3	95.7	96.8

确定特征融合方案和 f_{scale} 的值后,表 3-19 为 Market-1501 数据集的实际实验数据。实验数据表明部署此模块的确有效提高网络性能,并且两个不同的 f_{scale} 值所带来的网络性能提升非常接近。当 $f_{scale} = 0.5$ 时,网络的 R-1 性能指标较高,因此 Market-1501 数据集维持使用 $f_{scale} = 0.5$ 的特征融合参数。

表 3-19 Market-1501 数据集在不同 f_{scale} 下的性能指标

f_{scale}	mAP(%)	R-1 (%)	R-5 (%)	R-10(%)
0.5	87.9	94.8	98.3	99.1
0.7	87.9	94.7	98.5	99.2

此方案也在 Occluded Duke 数据集里进行实验验证。实验数据表明部署此模块的确有效提高遮挡行人重识别网络的性能。当 $f_{scale} = 0.5$,网络取得最好的性能。因此本文所提出的特征融合与重构模块的 f_{scale} 将固定在 0.5,以减少调参工作,降低研究人员的部署难度。

表 3-20 Occluded Duke 数据集在不同 f_{scale} 下的性能指标

f_{scale}	mAP(%)	R-1 (%)	R-5 (%)	R-10(%)
0.5	56.1	65.0	79.2	83.9
0.7	55.9	63.7	78.6	83.3

3.6 本章小结

本章通过网络框架图阐述在 TransReID 网络基础上新增的四个模块，组成本文提出的基于数据增强与特征融合的行人重识别网络。本文网络采用模块化设计，因此这些模块都可以轻易部署在其他现有的行人重识别网络。

本文新增的模块大致方向为数据增强与特征融合。基于数据增强的模块包括增强输入图像块，降低噪音的输入图像块全尺寸增强模块，有效提升网络效率的输入图像块随机屏蔽模块，和通过随机遮挡同一批次部分范围的特征序列，达成增强局部区域的关注特征学习目的的输出特征批次方块随机屏蔽模块。基于特征融合的模块为特征融合与重构模块，此模块加强输出特征序列中信息较多的输出特征块对网络的影响，减少背景噪声。

第四章 实验结果与分析

4.1 图表格式实验环境及设置

4.1.1 网络训练环境与配置

1. 服务器硬件配置

服务器 CPU 为两颗 Intel(R) Xeon(R) CPU E5-2630 v4, 频率为 2.20GHz。内存为 128GB, 显卡为 2 张 GTX1080Ti, 每张显卡均为 12GB 显存, CUDA 版本为 10.2。

2. 实验训练环境配置

实验均运行在 anaconda 虚拟环境, anaconda 版本为 4.10.3, 虚拟环境为 Python 3.6.11, 其中主要使用的库包括 PyTorch 1.6.0, torchvision 0.7.0 和 timm 0.3.2。

4.1.2 网络参数

1. 数据集预处理参数

训练集和测试集图像初始化大小均为 256×128 , 并归一化处理, 三通道均值和标准差均为 $[0.5, 0.5, 0.5]$ 。训练集图像所使用的数据增强方式包括随机剪裁、随机水平翻转(概率为 0.5), 随机擦除(RandomErasing, 概率为 0.5)^[48]和图像边缘填充(10 像素)。

2. 网络参数

网络训练期间, batch 大小为 64, 随机选取 16 个 ID 的行人, 每个 ID 均含有 4 张图像。网络使用 SGD 优化器, 动量系数为 0.9, 权重衰减系数为 $1e-4$ 。初始学习率为 0.008, 并采用余弦学习率衰减策略。网络将经过 120 回的迭代训练。主干网络为 ViT-B/16, 初始主干网络权重使用 ImageNet-21K 进行预训练, 并在 ImageNet-1K 数据集上进行微调。网络测试期间, batch 大小为 256。

4.1.3 行人重识别数据集概述

本文采用 3 个行人重识别数据集, 分别是 DukeMTMC-reID、Market-1501 和 Occluded Duke, 以客观评估网络性能与泛化能力。

DukeMTMC-reID 数据集是 DukeMTMC 数据集的子集, 是目前其中一个最被广为

使用的行人重识别数据集。研究人员通过杜克大学中 8 个不同的静态摄像机采集 85 分钟的高清视频, 构建 DukeMTMC 数据集, 并截取每 120 帧的视频画面, 获得 36411 张行人照片, 构建 DukeMTMC-reID 数据集。DukeMTMC-reID 数据集有 1404 个行人被超过 2 个不同的摄像机拍摄。408 个行人只被 1 个摄像机拍摄, 因此被作为干扰项。数据集随机采样 702 个行人, 共 16522 个行人边界框作为训练集, 并将剩余 702 个行人和干扰项, 共 17661 个行人边界框作为测试集。测试集中的行人在每个摄像机中随机选择 1 个行人边界框作为查询集, 共 2228 个。

Market-1501 数据集是研究人员通过 5 个高解析度摄像机和 1 个低解析度摄像机在清华大学采集的数据集。Market-1501 数据集含有 1501 个行人, 共 32668 张通过 Deformable Part Model (DPM) 行人检测器采集的行人边界框。通过 DPM 行人检测器采集的行人边界框会与人工标注的行人边界框做比较, 并将准确率较差的行人边界框作为干扰项。这类干扰项在数据集共有 2793 个。每个行人都被至少 2 个, 至多 6 个摄像机拍摄。数据集随机采样 750 个行人, 共 12936 个行人边界框作为训练集。剩余 751 个行人, 共 19732 个行人边界框作为测试集。测试集中的行人在每个摄像机中随机选择 1 个行人边界框作为查询集, 共 3368 个。查询集使用人工标注的行人边界框, 不同于训练集和测试集。

Occluded Duke 数据集是遮挡行人重识别数据集, 其行人的遮挡情况相比前两个数据集更复杂, 图像有各种大障碍物, 如自行车、骑车、树木和其他行人等。此数据集出自于 DukeMTMC-reID 数据集, 两者之间的差别为 Occluded Duke 数据集将 DukeMTMC-reID 数据集行人被遮挡的图像从测试集移到查询集。训练集共有 15618 个行人边界框, 测试集共有 16661 个行人边界框, 而查询集共有 2210 个行人边界框。

4.1.4 网络性能评价指标

为了更客观评估与比较不同行人重识别网络之间的性能差异, 本文采用累积匹配特征曲线 (Cumulative Matching Characteristic Curve, CMC Curve) 和全类平均正确率 (mean Average Precision, mAP) 两个指标, 两者都是重识别任务的热门评价指标。

重识别算法的目标是从测试集寻找与查询集图像最相似的图像。算法将通过提取查询集图像的特征, 计算查询集图像与测试集图像的相似度并将每次查询结果按相似度排序, 并由此从测试集寻找与之最相似的图像。Rank-k 表示按照相似度排序后前 k 张测试集图像中存在与查询集图像相同 ID 的准确率。本文使用 Rank-1、Rank-5 和 Rank-10 三个指标, 而累积匹配特征曲线是 Rank-k 的集合所绘制出的曲线, 反映不

同情况下的检索准确率。

查准率 (Precision) 和查全率 (Recall) 是了解全类平均正确率的先决知识。查准率是查询集同一 ID 的照片在查询结果中的比例, 而查全率是查询集同一 ID 照片出现在查询结果数量占总数的占比。尽管如此, 这两个指标忽略了查询结果的相似度排序。通过计算 P-R 曲线形成的面积, 就可以得出平均正确率 (Average Precision, AP), 综合了这张查询图像同 ID 结果的查准率和查全率。mAP 则是每个查询图像的 AP 求和后的平均值。

Rank-k 只考虑第一次命中的准确率, 然而实际上测试集内含有多个正确答案, 因此 Rank-k 无法评估靠后样本的命中概率, 而 mAP 考虑了所有在排序列表内的样本, 因此 mAP 能够反映检索行人在数据库中所有正确的图片排在排序列表前面的程度, 能够更加全面地衡量重识别算法的性能。

4.2 实验结果及分析

表 4-1 为基线网络与本文网络的性能指标比较表。除了基线, 所有网络均使用输入图像块随机屏蔽模块和输入图像块全尺寸增强模块, 并依序增加各个模块, 体现各个模块对网络的性能变化。表 4-1 中, PD 代表输入图像块随机屏蔽模块, PFDE 代表输入图像块全尺寸增强模块, BDB 代表输出特征批次方块随机屏蔽模块, FRM 代表特征融合与重构模块。

表 4-1 本文网络与基线网络的性能指标比较

网络	DukeMTMC-reID					Market-1501		Occluded Duke		
	PD	PFDE	BDB	FRM	mAP(%)	R-1 (%)	mAP(%)	R-1 (%)	mAP(%)	R-1 (%)
基线	-	-	-	-	80.3	89.8	87.6	94.4	56.0	64.0
网络 1	✓	✓	✓	-	81.0	89.8	87.7	94.2	57.1	64.3
网络 2	✓	✓	-	✓	80.9	89.9	87.7	94.2	57.4	64.7
网络 3	✓	✓	✓	✓	81.2	89.7	87.9	94.4	57.7	65.1

根据表 4-1 的数据, 多个模块的叠加使用并不会相互冲突, 共同提升网络性能。此外, 本文提出的行人重识别网络对 Occluded Duke 数据集效果最佳, 说明本文提出的网络确实可以从遮挡行人重识别数据集中提取更多具辨识度的特征, 达到提升网络

性能的目的。

4.3 消融实验

除了上述针对各个模块与完整网络的实验数据外，本文会针对网络的网络架构和其他不被采用的构思进行一系列实验，以寻找带来最高性能的网络。

4.3.1 网络架构研究

网络架构对行人重识别任务的性能起到非常关键的效果。此外，本文提出的各个模块的相互排斥性与先后顺序对网络性能的影响也是值得研究的重点。因此，不同网络结构之间的对比是重要的。本文提出两个网络架构设想：

方案 1：网络架构有四个输出分支。借鉴 TransReID 的网络架构，输出特征批次方块随机屏蔽模块、特征融合与重构模块和特征块重排序模块为并联结构，三者处理的输出特征序列分别进入不同分支并计算损失。

方案 2：网络架构依然保持两个输出分支。借鉴 PFT 的网络架构，经过特征融合模块后的输出特征序列将分为两个分支，第一个分支为不经过其他模块的全局分支；第二个分支中输出特征序列依次经过输出特征批次方块随机屏蔽模块和特征块重排序模块。

表 4-2 不同网络架构下的性能指标

	DukeMTMC-reID		Market-1501		Occluded Duke	
网络架构	mAP (%)	R-1 (%)	mAP (%)	R-1 (%)	mAP (%)	R-1 (%)
方案 1	81.2	89.7	87.9	94.4	57.7	65.1
方案 2	80.8	89.5	87.8	94.4	56.3	63.2

表 4-2 为不同数据集中不同网络架构方案下的性能指标。实验数据表明方案 1 的性能较方案 2 高。方案 1 能更好利用每个模块的特性，帮助网络从不同角度提取行人特征，从而提高网络性能。综上所述，本文采用方案 1 作为本文网络架构。

4.3.2 输出特征随机屏蔽模块

根据上述数据，输入图像块随机屏蔽模块确实提升网络效率。借鉴此想法，输出

特征序列是否可以经过相同模块，对特征序列进行随机屏蔽，以避免数据过拟合，提升网络性能？本文暂且称此模块为输出特征随机屏蔽模块 (Feature Patch Dropout)，其实现原理与输入图像块随机屏蔽完全相同。通过调整输出特征序列保留比例 (Feature Patch Keep Rate)，随机屏蔽不同数量的特征块。

下表为 DukeMTMC-reID 数据集在不同输出特征序列保留比例对网络性能的影响。此实验将同时使用输入图像块随机屏蔽模块，且没有运用输出特征批次方块随机屏蔽模块和特征融合与重构模块，步长为 16，图像解析度为 256×128 。根据实验数据，当 $rate_{keep}$ 为 1.0 时，对输出特征序列进行随机屏蔽并没有获得任何性能提升。相比之下，当 $rate_{keep}$ 为 0.7 时，对输出特征序列进行随机屏蔽都可以获得部分性能提升。

表 4-3 DukeMTMC-reID 数据集在不同 $rate_{keep}$ 和
输出特征序列保留比例下的性能指标与训练开销

$rate_{keep}$	输出特征序列保留比例	训练时间	mAP (%)	R-1 (%)	R-5 (%)	R-10 (%)	训练时间变化 (%)	性能变化 (%)	训练所用显存 (MiB)	训练显存变化 (%)
	1.0	3:57:51	80.3	89.8	95.6	96.8	—	—	5171	—
1.0	0.7	3:57:12	80.4	89.4	95.6	96.6	-0.27	+0.1	5543	+7.19
	0.5	3:56:37	80.3	89.4	95.4	96.7	-0.52	+0.0	5435	+5.11
	1.0	2:48:00	79.4	88.6	95.0	96.6	—	-0.9	4181	—
0.7	0.7	3:06:52	79.4	88.8	95.5	96.9	+11.23	+0.0	4593	+9.85
	0.5	3:08:52	79.5	88.8	95.1	96.7	+12.42	+0.1	4525	+8.23

尽管如此，本文最终并没有采用此方案，没有采用此方案的原因在于输出特征批次方块随机屏蔽模块。输出特征批次方块随机屏蔽模块不仅带来更大的性能提升，且相较于数据增强模块正则化数据或特征以防止网络过拟合，输出特征批次方块随机屏蔽模块可以增强局部特征。输出特征序列保留比例无法达成相同目标。最后，由于两者的部署位置的相同，同时使用两者会导致过多特征被移除，使性能下降，因此在权衡之下选择输出特征批次方块随机屏蔽模块。

4.4 本章小结

本文通过实验对比了基线网络与本文提出基于深度学习的行人重识别网络的性能。实验数据表明本文提出的模块确实提高网络性能，并且每个模块都发挥作用。此外，本章针对网络架构和其他不被采纳的模块进行消融实验，以寻找更好的网络。实验数据表明更多的输出分支让网络多方面提取行人特征，从而提升网络性能。由于多个模块之间的相互冲突，因此本文只选择使用带来最大效果的模块，从而放弃其余模块。

第五章 结论

本章将对本文工作进行总结，同时分析本文提出的方案对法律社会、环境和公共安全的影响，之后分析本文提出的方案存在的不足之处，提出本文网络优化的研究方向与可能方案。

5.1 本文主要工作

行人重识别任务是当前计算机视觉的热门研究方向，因遮挡、视觉变化、行人姿势等种种原因而成为具有挑战性的研究课题。目前其中一个研究方向为增加多个辅助信息和处理辅助信息的附加网络以帮助行人重识别网络做出更准确的判断，但辅助信息的存在可能导致行人重识别网络更复杂，且辅助信息的采集与人工标注耗时巨大。在这样的背景下，本文着重于不添加额外的辅助信息的前提，对现有的输入和输出进行数据增强与特征融合与重构，以提升性能与效率。本文采用模块化设计，在现有 TransReID 网络的基础上融合四个模块，详细工作总结如下：

5.1.1 基于数据增强的行人重识别网络

本文提出的数据增强模块共有三个，包括提升特征表示能力与降低噪音的输入图像块全尺寸增强模块，提升网络效率的输入图像块随机屏蔽模块和增强局部特征学习的输出特征批次方块随机屏蔽模块。

- 输入图像块全尺寸增强模块通过对输入图像块序列和相同大小的可学习张量进行哈达玛积，有效降低噪音，并凸显图像块中信息量较大的区域，进而提升性能。
- 输入图像块随机屏蔽模块通过随机屏蔽部分输入图像块，减少 ViT 主干网络的输入，达到提升网络效率的目的。
- 输出特征批次方块随机屏蔽模块通过在同一批次的特征序列中屏蔽相同位置的特征图，删除相同语义的身体部位，以增强局部区域的关注特征学习，增强网络的泛化能力与性能。

5.1.2 基于特征融合与重构的行人重识别网络

增强输出特征块的空间相关性的特征融合与重构模块通过将输出特征序列拆分

并对相对信息量较少的分支序列头乘上小于 1 的特征融合比例，再分别与两个主要特征序列相加，形成新的头尾序列，最后合并这些分支序列形成新的输出特征序列。此模块增强部分特征块的重要性，并减低噪声影响，以此增强网络性能。

5.2 非技术性分析

5.2.1 法律分析

随着行人重识别任务的深入研究，大规模的行人重识别数据集成为必备，研究人员必须通过架设更多摄像头在同一区域进行更长时间的数据采集。此外，随着行人重识别技术的大规模运用，运用此技术的个体或企业将掌握大量行人的移动轨迹。大规模的行人图像采集是否会造成资料外泄、侵犯他人隐私权？

根据《中华人民共和国个人信息保护法》，敏感个人信息是指泄露或非法使用可能导致人格尊严受到侵害或人身、财产安全受到危害的信息，其中包括行踪轨迹。处理敏感个人信息需要特定目的和必要性，并采取严格保护措施。在处理敏感个人信息时，必须获得个人的单独同意。对未满 14 周岁的未成年人个人信息的处理需要获得监护人的同意。个人信息处理者还应当向个人告知处理敏感个人信息的必要性以及对个人权益的影响。

由上述所知，行人重识别数据集收集与技术发展应当受上述法律规范。本文提出的行人重识别网络无需使用大规模行人重识别数据集也可以得到较高性能，因此才能在法律的规范下，安全、合法地发展并普及行人重识别技术。

5.2.2 公共安全分析

近期各国严加防范恐怖主义的滋生，使得确保大型公共场所的公共安全成为各国关注的治安问题。行人重识别技术与其他相关技术的配合，如人脸识别、属性识别等，可以检测各个行人的移动轨迹，迅速锁定嫌疑犯并实施抓捕行动。此技术的大力发展能有效提升执法效率，提升打击犯罪的效果。此外，行人重识别技术通过分析大型公共场所的人流量，提前在人流量大的区域部署更多维安人员，进一步降低犯罪率。从公共安全角度来说，行人重识别技术在未来依然具有相当大的研究潜能。

5.2.3 对环境与社会的影响

虽然基于视觉 Transformer 的行人重识别任务解决方案潜力巨大，但目前视觉

Transformer 所消耗的计算开销较大, 对能源开销要求也较高。此外, 目前人工智能的高速发展, 各大企业积极研发和训练人工智能, 对计算开销与能源开销要求日益增高, 这对于目前全球积极打造永续环境的目标相违背。本文提出的方案有效降低网络的能源成本, 网络更环保, 碳排放量的下降对全球环境可持续性有积极意义。

本文提出的行人重识别技术不仅可以运用在智能安防领域, 也可运用在大型公共场所的智能寻人、无人超市、智能机器人等领域。此外, 此技术可扩展到车辆重识别任务, 能实时检测交通车流量并优化交通路况。重识别技术的发展与应用有效降低人力资源的需求, 提高社会整体效率。尽管如此, 人力资源需求短期内的大幅下降有可能造成社会动荡。政府在之中有必要为受影响的人力资源提供经济和教育援助, 帮助他们适应新时代的就业环境。

此外, 分析数据的不当使用有可能威胁行人的的人身安全。部分人士基于对隐私权的疑虑反对利用监控视频数据进行分析。政府在这之中需扮演调解者的角色, 使用法律避免个人或企业肆无忌惮地非法收集与使用行人数据, 并且对行人数据进行加密, 以防数据外泄。

5.3 未来研究工作展望

本文主要对基于深度学习的行人重识别进行研究, 针对降低噪音、提升网络效率、增强局部特征学习和输出特征序列之间的空间相关性提出并融合相关模块。然而, 由于时间有限和行人重识别任务的复杂性, 本文还有需要解决的问题:

- (1) 本文在研究输入图像块随机屏蔽模块对客观评价网络效率上欠缺周全考虑, 并没有为每个网络记录计算开销, 仅考虑显存开销。尽管网络显存开销的下降也一定程度上表示网络所用计算开销的下降, 但本文确实没有对计算开销进行客观数据采集, 因此计算开销数据采集是未来研究工作之一。
- (2) 本文提及更换主干网络对效率提升有正面影响。本文原定在 TransReID 网络的基础上更换 DAT 主干网络^[44]。相比 ViT 主干网络, DAT 通过可变形自注意力机制, 对输入图像均匀布满参考点, 并将查询向量作为偏置网络输入计算对于参考点的偏置值。参考点偏移后通过双线性插值得到特征后, 借此推断出新的键向量和值向量。最后, 三个向量与通过偏移点计算的相对位置偏移作为多头注意力机制的输入。对行人重识别任务而言, 这项操作可以让网络将注意力集中在行人上, 图像信息含量较高的区域, 忽略背景噪音, 从而提升性能。由于更换主干网络工作

量巨大，因此最终并没有采取此方案。此方案可以和输入图像块随机屏蔽模块，共同提升网络性能和效率，可被作为未来研究工作之一。

- (3) 本文提出的特征融合与重构方案一直无法寻找合适的特征融合方案，以增强部分特征图的重要性，进而大幅增强网络性能。其他尚未在正文中提出的备选方案包括高斯权重方程，可学习特征融合比例、引入额外 ViT 层等。目前 ViT 的特征融合方案不如 CNN 成熟，因此适用于 ViT 的特征融合模块是未来的研究工作之一。
- (4) 为了更直观观察并比较各个网络对行人图像的特征提取差异，可视化工具必不可少。行人特征的可视化可以有效观察各个模块对网络的影响，并且可以分析行人特征可视化图寻找网络所忽略的行人特征，对症下药部署相对应的模块。由于时间关系，本文并没有实现各个网络行人特征可视化的比较，因此这方面是未来研究工作之一。

参 考 文 献

- [1] Zhu K, Guo H, Zhang S, et al. Aaformer: Auto-aligned transformer for person re-identification [EB/OL]. arXiv preprint arXiv:210400921, 2021.
- [2] Venkataramanan S, Ghodrati A, Asano Y M, et al. Skip-Attention: Improving Vision Transformers by Paying Less Attention [EB/OL]. arXiv preprint arXiv:230102240, 2023.
- [3] Chen X, Xu C, Cao Q, et al. Oh-former: Omni-relational high-order transformer for person re-identification [EB/OL]. arXiv preprint arXiv:210911159, 2021.
- [4] Pervaiz N, Fraz M, Shahzad M. Per-former: rethinking person re-identification using transformer augmented with self-attention and contextual mapping [J]. The Visual Computer, 2022: 1-16.
- [5] Li W, Zou C, Wang M, et al. DC-Former: Diverse and Compact Transformer for Person Re-Identification [EB/OL]. arXiv preprint arXiv:230214335, 2023.
- [6] Li Y, He J, Zhang T, et al. Diverse part discovery: Occluded person re-identification with part-aware transformer[C]// proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, F, 2021.
- [7] Lee K W, Jawade B, Mohan D, et al. Attribute De-biased Vision Transformer (AD-ViT) for Long-Term Person Re-identification[C]// proceedings of the 2022 18th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), F, 2022.
- [8] He S, Luo H, Wang P, et al. Transreid: Transformer-based object re-identification[C]// proceedings of the IEEE/CVF international conference on computer vision, F, 2021.
- [9] Zhao Y, Zhu S, Wang D, et al. Short range correlation transformer for occluded person re-identification [J]. Neural computing and applications, 2022, 34(20): 17633-45.
- [10] DeVries T, Taylor G W. Improved regularization of convolutional neural networks with cutout [EB/OL]. arXiv preprint arXiv:170804552, 2017.
- [11] Chhabra S, Venkateswara H, Li B. PatchSwap: A Regularization Technique for Vision Transformers [J]. 2022.
- [12] Li P, Li X, Long X. Fencemask: a data augmentation approach for pre-extracted image features [EB/OL]. arXiv preprint arXiv:200607877,

- 2020.
- [13] Zhang H, Cisse M, Dauphin Y N, et al. mixup: Beyond empirical risk minimization [EB/OL]. arXiv preprint arXiv:171009412, 2017.
 - [14] Singh K K, Yu H, Sarmasi A, et al. Hide-and-seek: A data augmentation technique for weakly-supervised localization and beyond [EB/OL]. arXiv preprint arXiv:181102545, 2018.
 - [15] Wang H, Du H, Zhao Y, et al. A comprehensive overview of person re-identification approaches [J]. Ieee Access, 2020, 8: 45556-83.
 - [16] Liu M, Zhao J, Zhou Y, et al. Survey for person re-identification based on coarse-to-fine feature learning [J]. Multimedia Tools and Applications, 2022, 81(15): 21939-73.
 - [17] Li W, Zhao R, Xiao T, et al. Deepreid: Deep filter pairing neural network for person re-identification; proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, F, 2014 [C].
 - [18] Zheng L, Shen L, Tian L, et al. Scalable person re-identification: A benchmark[C]// proceedings of the IEEE international conference on computer vision, F, 2015.
 - [19] Gou M, Wu Z, Rates-Borras A, et al. A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets [J]. IEEE transactions on pattern analysis and machine intelligence, 2018, 41(3): 523-36.
 - [20] Ristani E, Solera F, Zou R, et al. Performance measures and a data set for multi-target, multi-camera tracking[C]// proceedings of the Computer Vision - ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Part II, F, 2016. Springer.
 - [21] Wei L, Zhang S, Gao W, et al. Person transfer gan to bridge domain gap for person re-identification[C]// proceedings of the IEEE conference on computer vision and pattern recognition, F, 2018.
 - [22] Zhuo J, Chen Z, Lai J, et al. Occluded person re-identification[C]// proceedings of the 2018 IEEE International Conference on Multimedia and Expo (ICME), F, 2018.
 - [23] Zheng W-S, Li X, Xiang T, et al. Partial person re-identification[C]// proceedings of the IEEE international conference on computer vision, F, 2015.
 - [24] Li M, Zhu X, Gong S. Unsupervised person re-identification by deep learning tracklet association[C]// proceedings of the European conference on computer vision (ECCV), F, 2018.
 - [25] Miao J, Wu Y, Liu P, et al. Pose-guided feature alignment for occluded person re-identification[C]// proceedings of the IEEE/CVF

- international conference on computer vision, F, 2019.
- [26] 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, 44(6): 2872-93.
- [27] Zheng L, Zhang H, Sun S, et al. Person re-identification in the wild[C]// proceedings of the IEEE conference on computer vision and pattern recognition, F, 2017.
- [28] Li W, Zhu X, Gong S. Harmonious attention network for person re-identification[C]// proceedings of the IEEE conference on computer vision and pattern recognition, F, 2018.
- [29] Si J, Zhang H, Li C-G, et al. Dual attention matching network for context-aware feature sequence based person re-identification[C]// proceedings of the IEEE conference on computer vision and pattern recognition, F, 2018.
- [30] Guo J, Yuan Y, Huang L, et al. Beyond human parts: Dual part-aligned representations for person re-identification[C]// proceedings of the IEEE/CVF International Conference on Computer Vision, F, 2019.
- [31] Zhang Z, Lan C, Zeng W, et al. Densely semantically aligned person re-identification[C]// proceedings of the IEEE/CVF conference on computer vision and pattern recognition, F, 2019.
- [32] Sun Y, Zheng L, Yang Y, et al. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)[C]// proceedings of the European conference on computer vision (ECCV), F, 2018.
- [33] Lin J, Ren L, Lu J, et al. Consistent-aware deep learning for person re-identification in a camera network[C]// proceedings of the IEEE conference on computer vision and pattern recognition, F, 2017.
- [34] Wang J, Zhu X, Gong S, et al. Transferable joint attribute-identity deep learning for unsupervised person re-identification[C]// proceedings of the IEEE conference on computer vision and pattern recognition, F, 2018.
- [35] Zheng Z, Zheng L, Yang Y. Unlabeled samples generated by gan improve the person re-identification baseline in vitro[C]// proceedings of the IEEE international conference on computer vision, F, 2017.
- [36] McLaughlin N, Del Rincon J M, Miller P. Recurrent convolutional network for video-based person re-identification[C]// proceedings of the IEEE conference on computer vision and pattern recognition, F, 2016.
- [37] Chen D, Li H, Xiao T, et al. Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding[C]// proceedings of the IEEE conference on computer vision

- and pattern recognition, F, 2018.
- [38] Xu S, Cheng Y, Gu K, et al. Jointly attentive spatial-temporal pooling networks for video-based person re-identification[C]// proceedings of the IEEE international conference on computer vision, F, 2017.
 - [39] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [J]. Advances in neural information processing systems, 2017, 30.
 - [40] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]// proceedings of the IEEE conference on computer vision and pattern recognition, F, 2016.
 - [41] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale [EB/OL]. arXiv preprint arXiv:2010.11929, 2020.
 - [42] Devlin J, Chang M-W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding [EB/OL]. arXiv preprint arXiv:1810.04805, 2018.
 - [43] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]// proceedings of the IEEE/CVF international conference on computer vision, F, 2021.
 - [44] Xia Z, Pan X, Song S, et al. Vision transformer with deformable attention[C]// proceedings of the IEEE/CVF conference on computer vision and pattern recognition, F, 2022.
 - [45] Zhu X, Su W, Lu L, et al. Deformable detr: Deformable transformers for end-to-end object detection [EB/OL]. arXiv preprint arXiv:2010.04159, 2020.
 - [46] Liu Y, Matsoukas C, Strand F, et al. PatchDropout: Economizing Vision Transformers Using Patch Dropout[C]// proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, F, 2023.
 - [47] Dai Z, Chen M, Gu X, et al. Batch dropblock network for person re-identification and beyond[C]// proceedings of the IEEE/CVF international conference on computer vision, F, 2019.
 - [48] Zhong Z, Zheng L, Kang G, et al. Random erasing data augmentation[C]// proceedings of the AAAI conference on artificial intelligence, F, 2020.

致 谢

四年的本科求学生涯转瞬即逝，在学位论文工作的完结下即将划上句点。进行论文工作期间，我受到许多人的帮助与支持，此时此刻我谨向他们表示最衷心的感谢。

首先，我在此感谢我的导师。导师对相关课题的理解与经验对我进行论文设计工作提供莫大帮助。在论文选题、思路整理、论文工作进度跟踪和论文撰写与修改上，导师除了愿意抽出时间对我进行沟通之外，也给予我耐心的指导，让我受益匪浅。面对困难的当下，导师也愿意提出对于我的研究成果的看法，并指出正确方向。我在此向导师表示最真诚的感谢和敬意。

其次，我在此感谢我的师兄。他不仅愿意与我分享学习经验，也帮助我解决一些技术性问题。即使他之后顺利完成硕士学位并离开校园，他也愿意抽出时间为我解答种种疑问。在我担心学位论文进度的当下，他也给予支持与鼓励，激励着我继续向前迈进完成学位论文。

我在此也要感谢同是自动化系的同学与伙伴。与他们培养坚毅友情的同时，大家一同进行学位论文工作时不仅相互提醒论文工作进度也相互沟通获取更多灵感，让我更有动力完成学位论文工作。此外，我也必须感谢我的家人，为我提供无条件的支持与爱，能让我在异国他乡完成学位。他们的鼓励、支持与期望是我持续努力完成学位工作的动力之一。

最后，再次感谢所有给予我帮助、支持与鼓励的人，让我能够顺利完成这篇学位论文。我将永远铭记你们的恩情和帮助，为我所在的团队和社会做出更多的贡献。

DEEP LEARNING-BASED PERSON RE-IDENTIFICATION

Person re-identification(Person ReID) is one of the hot research topic in computer vision. The main research purpose is to accurately identify the same person across multiple non-overlapping surveillance cameras with different viewpoints. The research is significant because it has wide real world application such as video surveillance, security detection, and crowd management.

Person ReID has encountered various challenges such as occlusions, illumination changes, viewpoint changes and complex background. Currently, feature representation learning and metric learning are used to determine features from the person images. Researchers also studied and applied deep learning-based methods to address these challenges. Convolutional neural networks(CNNs) are been used as backbone networks to extract features from images.

However, there are a few limitations from CNNs. CNNs mainly focus on small discriminative regions. Besides, since CNNs use convolution and downsampling, the information of images is not retained completely. Thus, the usage of Vision Transformer(ViT) as backbone network is becoming increasingly popular to address these challenges. ViTs have stronger sequence information capturing ability due to the usage of self attention. Despite of these advantages compared to CNNs, ViTs still suffer from high computational complexity and low efficiency. Researchers have proposed various ViT variants to counter this issue and adapt to person ReID, such as AAformer, SKIPAT, OH-Former and Per-former.

In addition, as global and local features alone cannot fully differentiate the differences between two persons, person ReID researchers has gradually shifted the direction towards the integration of auxiliary information. However, collecting and organizing auxiliary information is complex and increases the complexity of the network. TransReID and PFT provides different perspective to address the challenges without complex auxiliary information and related network. Therefore, relying on existing feature maps for more detailed data augmentation and feature fusion, as well as better utilizing existing data and feature information, is also an important research direction.

In this paper, TransReID is used as baseline, which is the first pure ViT approach to solve person ReID challenges. TransReID do have some innovation to enhance the feature extraction ability. First of all, TransReID uses overlapping patches to retain local adjacent structure. Side Information Embedding(SIE) encodes camera ID and viewpoint ID information, incorporates non-visual information in patch sequence. Jigsaw Patch Module(JPM) is proposed to rearrange patch sequence by shift and shuffle operations, expands long-range dependencies and extracts more discriminative, robust features.

To address these issues, this paper proposes a person ReID network based on TransReID. This paper improves the efficiency and performance based on data augmentation and feature fusion & reconstruction, which includes four modules. Among them, the data augmentation-based person ReID network contains modules below, which are Patch Full-Dimension Enhancement (PFDE) module, Patch Dropout module and Batch DropBlock(BDB) module.

PFDE module is a simple yet effective module to enhance feature representation. Besides, this module is easy to deploy to any ViT. First, the module takes patch sequence without [cls] token, positional embeddings and side information embeddings as input. The size of patch sequence is where N is the number of patches and D is the dimension of each patch. A learnable tensor with same size as the patch sequence is initialized with all elements equal to 1. The Hadamard Product between patch sequence and learnable tensor is calculated to obtain enhanced patch sequence.

Patch Full-Dimension Enhancement

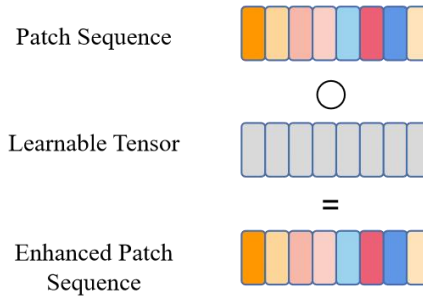


Fig 1: Module Structure of Patch Full-Dimension Enhancement Module

The element value of the learnable tensor will be decreased if there are less information or noises in the patch, thus the impact of these patches to the network will be

decreased. While the element value of the learnable tensor will be increased if there are more information in the patch, the impact of these patches to the network will be increased. This module successfully improves performance without adding more training time, computational and memory cost.

Patch Dropout module is used to increase the efficiency of ViTs without changing to other ViT variations, thus successfully avoids high workload to change the backbone networks. Since not all patches are equally important, the number of patch in patch sequence can be decreased, so that the input is decreased, thus improves model efficiency.

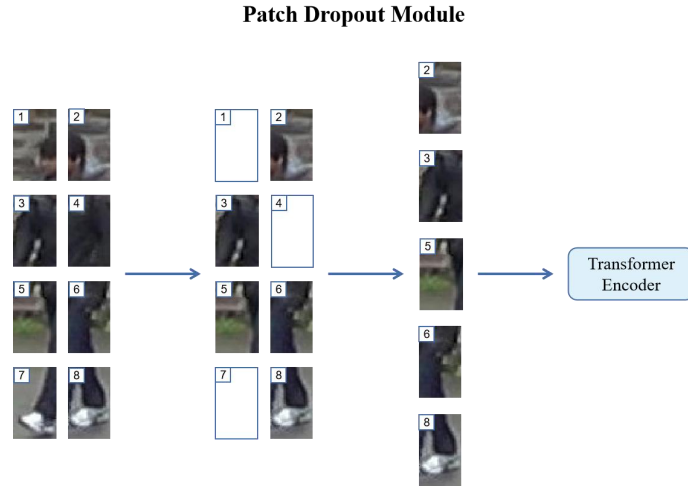


Fig 2: Module Structure of Patch Dropout Module

Patch Dropout module takes patch sequence, including [cls] token, positional embeddings and other auxiliary information as input. [cls] token will be retained by default. The module randomly retains certain amount of patch, according to the patch keep rate. According to the experiment result, this module has significantly decreased the computational and memory cost, while only caused minor performance degradation.

BDB module is designed to enhance the network ability to extract the attentive feature of local regions. However, BDB module is only available to CNNs. Therefore, we propose BDB module that can be applied to ViT.

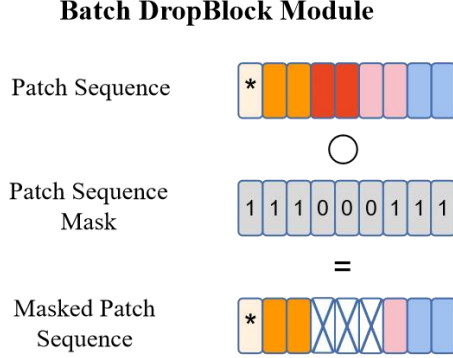


Fig 3: Module Structure of Batch DropBlock Module

BDB module takes output of 11th ViT layer, patch drop rate ($ratio_h$), number of patches in width ($patch_x$) and number of patches in height ($patch_y$) as input of module. Firstly, the number of patches in height that needed to be masked, $mask_y$, is the multiplication between $patch_y$ and $ratio_h$. The number of patches in width that needed to be masked, $mask_x$, is same as $patch_x$. Therefore, the size of the patch mask is the multiplication between $mask_x$ and $mask_y$.

A random number will be chosen between 0 and $(patch_y - mask_y - 1)$ to determine the starting row of the patch mask, which is $mask_{start}$. Therefore, patches starting from $(mask_{start} \times patch_x)$ to $(mask_{start} + mask_y) \times patch_x$ will be masked. [cls] token will not be masked so that the global feature extracted by backbone network is retained. After the mask is formed, the Hadamard Product between patch sequence and mask is calculated to obtain the enhanced patch sequence.

At the same time, the feature fusion and reconstruction-based person ReID network contains module below, which is Fusion & Reconstruction Module(FRM). FRM module is used to enhance the spatial correlation of the output feature block, reduce noise impact and improves model performance by enhancing the impact of more significant patch sequence.

There are more important regions and less important regions within an image. According to the calculation result of cosine similarity between local features blocks and global feature, more important regions within an image are mostly located in the middle of patch sequence. Besides, the front and end of the patch sequence usually have noises from background or other items, thus may cause disturbance to the model. However, it doesn't mean that we should get rid of these less important regions since there are still some

information in these regions that might can be used to determine different person.

Fusion and Reconstruction Module

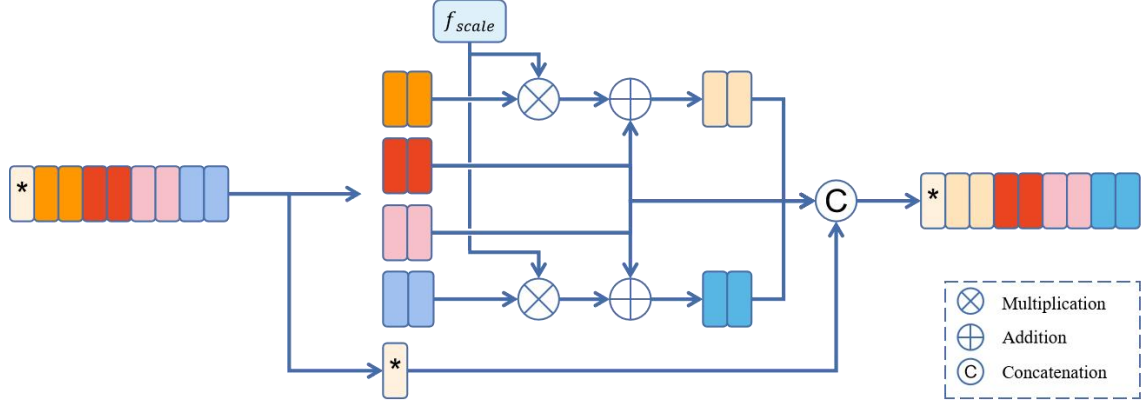


Fig 4: Module Structure of Fusion and Reconstruction Module

First, FRM splits the [cls] token and patch sequence so that [cls] token is not affected. The patch sequence will be split into four different sequences, which are head sequence, second sequence, third sequence and tail sequence. To decrease the impact of less important regions to the output, head sequence and tail sequence will firstly multiplied by feature fusion scale, f_{scale} , which is smaller than 1. Head sequence and tail sequence will fuse with second patch sequence and third patch sequence respectively. These sequences and [cls] token will be concatenated to form new patch sequence.

In conclusion, the proposed person ReID network in this paper has good practicality and application prospects. Compared with traditional convolutional neural networks, this network effectively improves the performance and efficiency of person ReID tasks through methods such as data augmentation and feature fusion reconstruction.