SwinCascador:面向行人搜索的端到端联合优化框架

刘玉杰　柴世杰　景建森　张明珠　李宗民
(中国石油大学(华东)青岛软件学院、计算机科学与技术学院　山东 青岛 266580)

摘　要　在全景图像中的行人定位和匹配任务中,面临着平衡检测和重识别任务的挑战,检测关注行人的共性特征,而重识别侧重区分个体差异。为联合优化两个子任务,提出了一种基于重采样机制的端到端多阶段行人搜索框架。采用增强头部网络和粗到细的筛选策略,实现两个任务的协同优化。此外,通过解耦并分离独特性特征加以角度分量监督,增强了同一身份与不同身份之间的区分能力,并通过边界约束降低了身份冲突的风险。在 CUHK-SYSU 和 PRW 基准数据集上的实验结果表明,该方法的 mAP 分别达到了 95.15%和 58.29%,top-1 准确率分别达到了 95.79%和 87.21%,本方法在两种数据集的精度表现均优于现有方法。

关键词　行人搜索　端到端　Swin transformer　角度增强归一化

中图分类号　TP391　　**文献标志码**　A　　DOI: 10.3969/j.issn.1000-386x.2024.12.001

SWINCASCADOR: AN END-TO-END JOINT OPTIMIZATION FRAMEWORK FOR PERSON SEARCH

Liu Yujie　Chai Shijie　Jing Jiansen　Zhang Mingzhu　Li Zongmin
(*Qingdao Institute of Software, College of Computer Science and Technology, China University of Petroleum (East China), Qingdao 266580, Shandong, China*)

Abstract　Pedestrian localization and matching in panoramic images face the challenge of balancing detection and re-identification tasks. Detection focuses on the common features of pedestrians, while re-identification aims to distinguish individual differences. To jointly optimize these two sub-tasks, we propose an end-to-end, multi-stage pedestrian search framework based on a resampling mechanism. The framework employs an enhanced network head and a coarse-to-fine filtering strategy to achieve collaborative optimization of both tasks. Additionally, by decoupling unique features with angular component supervision, the framework enhances the ability to distinguish between different identities and reduces the risk of identity conflict through boundary constraints. Experimental results on the CUHK-SYSU and PRW datasets show that our method achieves mAPs of 95.15% and 58.29%, with top-1 accuracies of 95.79% and 87.21%, respectively, demonstrating superior performance over existing methods on both datasets.

Keywords　Person search　End-to-end　Swin transformer　Angle enhanced normalization

0　引　言

行人搜索技术在社会公共安全中具有重要作用,特别是在视频监控系统中,它能够准确地识别并查询目标行人。通过在全景图像中快速地定位特定行人,从而在多摄像头监控环境中提高安全性和效率。行人搜索不仅仅是单一的目标检测任务,它同时需要结合行人检测和重识别技术,因此如何提升行人搜索的准确性,成为了一个重要的研究课题。行人搜索通过多任务学习的方式,耦合行人检测与重识别任务,旨在针对同一目标行人进行搜索。在行人重识别任务中,系统主要关注识别和区分不同个体的特征;而在行人搜索中,除了重识别外,检测任务还需要解决如何从原始图像中裁剪并调整目标行人的问题。特别是在没有位置和比例先验知识的情况下,行人搜索需要有效地从大范围的全景图像中查找并匹配目标行人。

行人搜索通常由行人检测与匹配过程组成,现有的方法大致可分为两种:一种是独立考虑两个任务的两阶段模型(如图 1(a)所示),另一种是将多个任务联

收稿日期:2024-12-22。国家重点研发计划(2019YFF0301800);国家自然科学基金(61379106);山东省自然科学基金(ZR2013FM036,ZR2015FM011)。**刘玉杰**,副教授,主研领域:图形图像处理,计算机视觉。**柴世杰**,硕士研究生。**景建森**,硕士研究生。**张明珠**,硕士研究生。**李宗民**,教授。

合的端到端模型(如图 1(b)所示)。在两阶段模型[1-3]中,检测任务侧重于学习泛化特征以捕捉更多行人,而重识别任务[4]则注重提高个体差异的区分能力。由于检测优先和重识别优先的任务冲突,依赖单一策略难以平衡这两个子任务的优先级,从而影响搜索精度。

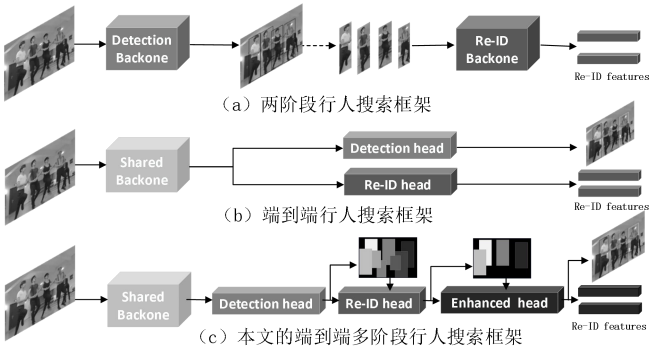

图 1 端到端多阶段网络架构与其他方法比较

端到端方法的目标是消除阶段划分,通过从输入图像中直接检索目标行人,简化处理过程。将检测与特征提取无缝衔接,使得网络可以同时学习检测与特征匹配。Xiao[5]等提出第一个端到端的行人搜索框架,Li[6]等设计了上下文双向图匹配的后处理算法,有效利用上下文信息作为行人匹配的重要补充线索。Yan[7]等提出了一种 anchor-free 的检测方法,直接检测和识别行人关键点。Gaikwad[8]等提出了一种基于隐私保护行人搜索模型,以提升在不同拥挤场景下的搜索能力。Zhang[9]等提出上下文实例批次采样方法,以增强细粒度特征表示和利用上下文信息构建训练批次。然而,由于检测与重识别并行进行,无法确保在特征提取前获得精确的边界框。这导致共享特征可能包含错误信息,尤其是低质量的提议对细粒度的重识别任务造成显著影响。尽管这些低质量特征对粗粒度的分类影响较小,但对精确匹配的重识别任务则有较大的负面作用。

当前的端到端方法引入 Vision Transformer[10](ViT)架构,将行人图像区域转化为序列数据,来提升行人搜索模型的性能。Cao[11]等通过引入遮挡注意力机制,进一步调整了模型的姿态与尺度不变性。而 Yu[12]等则提出了专门的 PSS 模块,结合多层次监督和部分注意力机制提升了检测与辨别性特征的学习。然而,ViT 捕获的细粒度特征是固定尺寸的,这使其在处理行人搜索这种尺度变化明显的任务时表现不佳。本文面临两个主要挑战:(1)共性与独特性目标的平衡,共性目标主要关注如何在不同图像中识别并定位更多的行人,而独特性目标则专注于提取每个行人的独特特征,实现有效的重识别。两者的冲突体现在:如何在保证检测准确性的同时提取足够的共性特征,以及如何确保足够的区分性特征,以区分不同身份的行人。现有方法在协调这两个目标时常常难以取得平衡,导致精度较低。(2)行人搜索任务需要同时保证同一身份内特征的一致性,并增强对不同身份之间差异的区分能力。然而,传统方法虽然在背景与人物的区分上有所成效,但在高相似度的身份区分方面表现不足,未能有效提取具有高辨识度的特征,从而限制了重识别的准确性。

为解决共性与独特性目标平衡问题,本文引入了 Swin Transformer[13]。其分层设计与移动窗口机制能够适应行人搜索中目标尺度的变化,本文提出端到端的多阶段级联策略(如图 1(c)所示),逐阶段细化检测结果,显著提升了检测器的鲁棒性,尤其在处理尺度变化大的目标时效果明显。借鉴 Cascade R-CNN[14]的架构思想,本文提出了三阶段级联检测策略,从粗到精逐步优化检测与重识别任务,缓解了二者之间的潜在冲突。在初始阶段,重点是准确检测目标的存在,并生成更多的候选框;后续阶段进一步细化候选框,提高检测精度。在第二和第三阶段,还引入身份匹配任务,深入挖掘个体差异,进一步提升重识别性能。

为解决身份特征的一致性与差异性问题,本文设计了角度增强归一化(AEN)损失函数。通过引入角度信息与余弦权重,AEN 损失能够显著提升特征的区分度。通过范数方法,利用特征向量的长度来区分背景与行人,并在同一特征空间内进行解耦。结合角度信息,提高对具有相似范数但方向不同的特征向量的区分度,减少误识别风险。余弦权重进一步增强了对特征向量方向差异的辨别能力,并通过合理设置角度边界,避免了特征向量的过度重叠,从而提升了识别精度。

本文提出的 SwinCascador 框架通过多阶段级联策略和 AEN 损失函数的结合,有效解决了行人搜索任务中的共性与独特性目标的平衡问题,以及身份特征一致性与差异性的区分问题。本文的主要贡献包括:

(1)提出了基于 Swin Transformer 的多阶段端到端行人搜索框架,能够有效适配多尺度目标并缓解检测与重识别任务的冲突。

(2)设计了 AEN 损失函数,通过引入角度信息与余弦权重,显著提高了特征区分能力。

(3)提出从粗到精的级联检测与重识别策略,逐步优化候选框与身份特征,为行人搜索任务提供了一种鲁棒高效的解决方案。

1 端到端行人搜索联合优化框架

1.1 重采样筛选机制

本文提出了一种重采样筛选机制,通过级联回归策略逐阶段提高候选框(proposal)的 IoU 阈值,从而有效缓

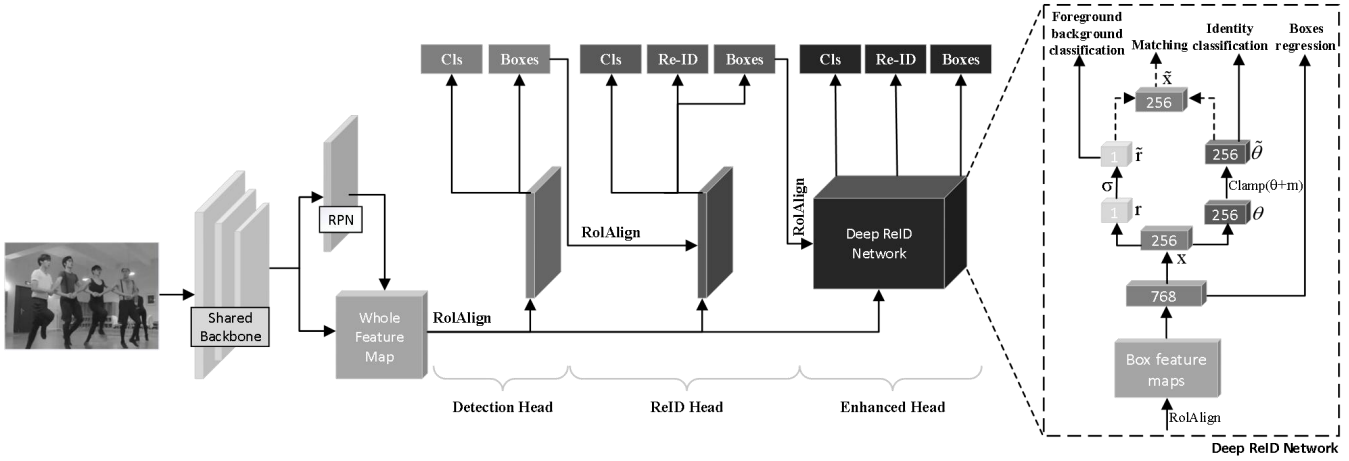

图 2 端到端多阶段网络架构

解行人检测与重识别任务之间的性能冲突。在传统方法中,候选框的质量通常不均,且阈值设定不当可能导致难样本被过早过滤,影响检测精度。因此,本文提出的筛选机制在训练过程中逐阶段调整 IoU 阈值,避免了由于阈值过高或过低导致难样本被过滤掉的风险,确保了每个阶段都能根据实际情况优化候选框的质量。

具体来说,在训练的初期阶段,由于候选框质量较低,IoU 阈值设置较低,可以保留更多的候选框以涵盖更广泛的样本分布,防止难样本在特征表达不充分的情况下被筛选。保留更多候选框随着训练的深入,每阶段 IoU 阈值增加 0.1,使得模型能够逐步淘汰低质量的候选框,仅保留那些与目标高度匹配的高质量框。这一阈值调整策略基于实验数据的分析与理论假设:初期较低的阈值能够涵盖不同质量水平的候选框,便于模型学习广泛分布的样本特性,而后续逐步提高阈值能够有效减少冗余,聚焦于高质量候选框,从而在性能与效率之间取得最佳平衡。有效避免了在早期阶段因阈值过高而导致难样本被误剔除的问题,同时也避免了阈值过低带来的过多冗余框干扰,从而确保每个阶段的候选框质量不断优化。

通过这种逐阶段筛选和优化机制,重采样过程能够有效提高候选框的位置和尺寸精度,同时减轻检测任务中的过拟合问题,提升模型的泛化能力。此外,该机制为后续的行人重识别特征提取提供了更可靠的候选框基础,使得行人重识别任务能够在高质量的候选区域上进行,从而显著提高了匹配的准确性和稳定性。

1.2 增强头部模块

该模块旨在通过设计深度重识别网络实现行人检测与重识别任务的高效平衡。如图 2 框线所示,该模块首先通过 RoI Align 提取候选区域的特征图(Box Feature Maps),生成 768 维的特征嵌入,并输入深度重识别网络进行前景/背景分类、身份匹配分类和边界框回归三项任务的并行处理。其中,前景/背景分类用于判定目标区域是否为行人,增强对目标与非目标区域的区分能力,身份匹配分类通过计算特征嵌入与身份标签的交叉熵损失进一步区分不同的行人身份,边界框回归则负责精确修正目标检测框的位置。在特征嵌入处理中加入多个 256 维全连接层进行处理,以逐步优化特征表达,同时通过非线性变换增强对身份特征的区分能力。

这种设计通过共享嵌入特征实现多任务联合优化,既有效避免了检测与重识别任务间的潜在冲突,又充分利用了特征信息的互补性,显著提高了系统整体性能。特征的分块处理保证了各个任务能够专注于自身目标,减少特征冗余,确保了模块在复杂背景和目标密集场景中的鲁棒性,为行人搜索提供了更加精准可靠的解决方案。

1.3 SwinCascador 端到端多阶段框架

SwinCascador 是一种基于 Swin Transformer[13] 的端到端多阶段框架,旨在解决行人搜索任务中的挑战。通过结合重采样筛选机制和增强头部模块,框架显著提高了检测和重识别的性能,并提升了模型在复杂场景下对多尺度目标的识别能力。SwinCascador 包含以下几个关键部分:多阶段优化策略:框架从粗到细的逐步优化检测和重识别任务,初始阶段重点关注候选框的广泛性与覆盖率,通过 RPN 生成多样化的候选区域,后续阶段则逐步优化候选区域的精度与特征表达能力,最大限度地减少了二者之间的冲突。多尺度特征提取:通过提升多尺度特征提取能力,框架在特征提取过程中动态调整特征分辨率与感受野大小,以提升对小尺度目标与复杂背景的适应性,增强了对尺度变化大的目标的识别能力。级联检测策略:采用多阶段检测策略,逐步优化候选框的空间定位与分类能力,不断细化候选区域,提高候选框质量,为后续的特征提取和匹配提供保障。

具体来说,SwinCascador 构建了一个多阶段级联结构,用于高效学习行人检测与重识别的特征。在第一阶段,利用 Swin Transformer 和 RPN(区域提议网络)生成

候选区域,并通过检测头对这些候选区域进行分类和回归,从而得到初步的检测结果,为后续阶段提供高质量的初始候选框。第二阶段通过重识别头提取检测框中的特征,并进行身份匹配,进一步细化候选区域,提高检测的准确性。第三阶段引入增强头模块,通过深度特征优化与任务分工处理显著提升检测框的精度与身份匹配的稳定性。通过这种逐阶段的优化设计,检测与重识别任务的性能得到了显著提升,为行人搜索提供了更加精确、可靠的特征表达。

1.4 基于角度分量的强化 AEN 损失

基于范数与角度解耦的思想,提出了一种用于行人搜索任务的特征优化方法。传统的范数方法虽然能够区分背景和人物,但在处理不同身份行人相似性时,难以提取具有高区分度的特征。而行人搜索任务需要在确保同一身份内特征一致性的同时,增强对不同身份间差异的辨别能力。为此,本文通过将特征向量的范数与角度解耦,利用范数作为检测置信度的衡量依据,角度则用于行人重识别中的余弦相似度计算,增强对不同身份间的区分度。

通过在极坐标系中的显式分解,将 x 拆分为范数 r 和角度 θ 。

$$\mathrm{x} = \mathrm{r} \cdot \theta \tag{1}$$

式中:引入非线性激活函数 σ ,对输入特征 r 进行归一化处理,$\tilde{\mathrm{r}}$ 代表归一化后的输出特征,$E[r]$ 为 r 的均值,用于消除输入特征的偏移,$Var[r]$ 是 r 的方差,用于防止分母为零引发数值不稳定问题。通过 γ 和 β 对标准化后的特征进行线性变换。

$$\tilde{\mathrm{r}} = \sigma\left(\frac{r - E[r]}{\sqrt{Var[r] + \varepsilon}} \cdot \gamma + \beta\right) \tag{2}$$

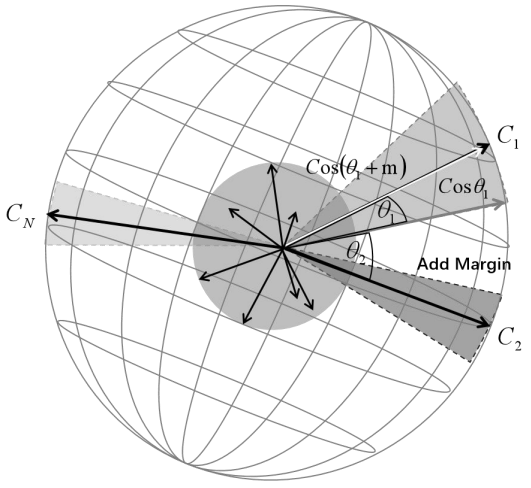

图 3 AEN 损失

然而,仅仅通过区分相似外观来处理不同类身份并不够。为增强对不同身份间差异的敏感性,本文提出了一种监督项 AEN,用于在特征空间中实现行人与背景的去中心化表示。通过范数长度,将同一特征空间中的背景与行人区分开来。图 3 中黑色短箭头表示背景,不同颜色区域代表不同身份,灰色平面为决策边界,黑色长箭头指示特征中心,虚线表示边缘特征。此外,在角度 θ 上引入角度余量 m,使得同类特征通过角度约束聚集,而不同类特征之间的差异被放大。

式中:角度余量 m_{ij} 为形状为[批量大小,6000]的矩阵,矩阵的每个元素 a_{ij} 表示样本的特征 i 与类别 j 的特征之间的角度余量。

$$m_{ij} = \cos\begin{pmatrix} a_{11} & \cdots & a_{1j} \\ \vdots & \ddots & \vdots \\ a_{i1} & \cdots & a_{ij} \end{pmatrix}, \quad if \quad j = label_i \tag{3}$$

通过加法角度惩罚机制,增强了类内特征的紧凑性,同时扩大了类间特征的间距。此外,OIM[5]损失用于标记和未标记身份的训练,本文将其损失值通过标签填充矩阵 m 进行计算:

$$\mathcal{L}_{OIM} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,j} \cdot \cos(\theta_{ij} + m) \tag{4}$$

通过调整中心线向量和边缘特征向量 x 的角度来增强特征向量之间的相似性。式中: N 是样本的数量,C 是类别的数量,角度是余弦矩阵 m 处理的角度。 y 是二进制指示函数,其中当样本 i 属于类别 c 时,它等于 1,否则它等于 0。下标 i 和 c 表示矩阵的第 i 行和第 c 列中的元素,指示样本 i 和类 c 之间的余弦相似性。

$$\mathcal{L}_{AEN} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c} \log\left(\frac{e^{\cos(\theta_{ic}+m)}}{\Sigma_{j=1}^{C} e^{\cos(\theta_{ij})}}\right) \tag{5}$$

结合 OIM 损失和 AEN 监督形成第三阶段的 AOIM 损失,其中 λ_{OIM} 和 λ_{AEN} 是平衡参数:

$$\mathcal{L}_{AOIM} = \lambda_{OIM}\mathcal{L}_{OIM} + \lambda_{AEN}\mathcal{L}_{AEN} \tag{6}$$

式中: p_i 表示行人特征 x 属于特定类别的概率,L 表示查找表,Q 代表存储特征的循环队列, v_i^T 和 u_k^T 分别为查找表和循环队列的存储特征, τ 为用于调整分布温度系数。

$$p_i = \frac{\exp\left(v_i^T x/\tau\right)}{\sum_{j=1}^{L}\exp\left(v_i^T x/\tau\right) + \sum_{k=1}^{Q}\exp\left(u_k^T x/\tau\right)} \tag{7}$$

通过最大化目标标签的预测概率 ,从而最小化重识别损失。

$$\mathcal{L}_{\mathrm{Re}ID} = \mathrm{E}x[\log pt] \tag{8}$$

本文的总损失函数将 AOIM 与其他模块联合优化:

$$\mathcal{L}_{ALL}=\sum_{t=1}^{T}\mathcal{L}_{\det}^{t}+\mathcal{L}_{ReID}+\mathcal{L}_{AOIM} \tag{9}$$

检测损失的设计与 SeqNet[15]中的形式完全一致,但采用了三个阶段叠加的方式,其中阶段数量默认为3。这一设计通过检测、重识别和增强头模块的端到端联合优化,有效提升了模型的整体性能。

2 实验及结果分析

2.1 数据集设置

CUHK-SYSU 是一个大规模的行人搜索数据集,以应对真实世界中的各种挑战,如视角,光照和背景的变化,涵盖城市场景和电影快照,共包含 18184 幅图像。其中有 8432 个标记的身份和 96143 个标记的边界框。本文在实验中使用默认图库大小为 100 来验证结果。

PRW 相较于 CUHK-SYSU 更具挑战性,数据集中包含摄像头位置信息,且聚焦于大学校园内的行人搜索任务。该数据集由 6 台摄像机拍摄,视角多样。此外,标注密度较低,在 10 小时的视频中,每隔 25 帧才标注一次。这种稀疏的标注数据,使得在非重叠视角和时间间隔较大的摄像头之间匹配行人变得更加困难,从而增加了行人搜索任务的复杂性。

2.2 参数设置

本文在 PyTorch 框架中实现了方法,实验在 Tesla P100 GPU 上进行。采用 Swin Transformer[11]的微型版本(swin-T)作为骨干网络。输入大小被调整为 1504×928,由于实验环境的显存受限,使用半精度方法降低显存的使用。对于 PRW 数据集,模型进行了 18 个 epoch 的训练,batchsize 为 3;对于 CUHK-SYSU 数据集,模型进行了 20 个 epoch 的训练,batchsize 为 2。权重衰减和动量分别设为 5×10^(-4)和 0.9。学习率设置为 0.003,对于 CUHK-SYSU/PRW,OIM 的循环队列大小设置为 5000/500。

2.3 训练和推理

在训练阶段,本文的网络进行了端到端的行人检测和重识别训练。行人检测损失由三个阶段的回归和分类损失项组成,每个阶段根据不同的阈值进行微调。实验结果如表2所示,表明在三个阶段中,使用阈值分别为0.5、0.6 和 0.7 时,达到了最佳性能。级联回归会在连续的阶段中对提议进行重新采样,以适应更高的 IoU 阈值。为了增强同一身份内部特征的相似性,并降低与其他身份的相似性,引入了一个角度阈值 m,初始设定为0.5。角度增强损失的平衡系数针对不同数据集进行了调整:对于 PRW 数据集为 0.02,对于 SYSU 数据集为0.04。

在推理阶段,RPN 网络提出了 300 个提议框,这些提议框直接进入 Roi Pooling,进行类别分类和框的回归。与训练阶段不同的是,在推理阶段无法对这些提议框进行采样,因为推理阶段无法获取真值信息,也因此无法计算 IoU。

使用累积匹配特征(CMC)和平均精度(mAP)来评估算法的效果。CMC 用于衡量从图库中查找查询图像时的 top-K 准确率。而 mAP 则通过计算精确率-召回率曲线,获得每个查询的平均精度,然后对所有查询的平均精度进行整体平均。这两个指标提供了全面的性能评估。

2.4 消融实验

本文对 PRW 数据集进行了一系列消融研究,以分析本文的设计决策。为验证各方法组合的有效性,本文提供了使用不同部分组合的消融实验结果,实验结果如表 1 所示。本文的基线模型采用了 tiny 规模的预训练模型作为先验知识。在表 1(a)中,加入 AEN 角度监督损失函数后,mAP 从 56.84%提高到 57.47%,top-1 精度从86.78%提高到 87.07%。表 1(b)中加入增强头组成多阶段结构显著提高了行人搜索的精度。表 1(c)展示了两种方法叠加起来取得更好的效果。

表 1 每个组件的消融研究(%)

实验	AEN	增强头	mAP	top-1
基线			56.84	86.78
(a)	√		57.47	87.07
(b)		√	57.30	86.63
(c)	√	√	**58.29**	**87.21**

为证明端到端多阶段结构的贡献,本文通过根据级联的不同阶段改变 IoU 阈值大小来从粗到细评估约束的效果。实验结果如表 2 所示,逐步增加级联结构的阈值显著提高了行人搜索的准确率,mAP 从 56.89%增加到 58.29%。通过实验可知,当阈值过低时将引入更多的相似身份,增加后续行人匹配的难度。当阈值太高时,较高的阈值导致在先前阶段中过滤掉真实值,从而降低top-1。

表 2 调整三个阶段的 IoU 阈值(%)

检测头	重识别头	增强头	mAP	top-1
0.4	0.4	0.4	56.89	86.53
0.5	0.5	0.5	57.57	86.24
0.6	0.6	0.6	57.06	86.00
0.5	0.6	0.7	**58.29**	**87.21**

表 3 检测为真值时的精度(%)

方法	检测		重识别	
	mAP	top-1	mAP	top-1
SeqNet[15]	100	100	47.9	85.1
COAT[12]	100	100	54.7	**88.0**
本文	100	100	**61.57**	87.85

为验证对行人匹配阶段改进的有效性,本文通过选择用真值检测框来消除检测阶段的干扰。实验结果如表3所示,当检测为真值时,行人匹配部分具有更高的准确率,实验结果表明本文方法在精度方面有显著提高。

本文探究了平衡系数λREN对模型性能的影响,实验结果如图4所示,λ_{REN}在0.020时Top-1准确率最高,而在0.050时mAP表现最好。这表明平衡系数对不同性能指标的优化效果不同。为了进一步提升模型的鲁棒性和整体性能,本文探索更精细的超参数调优策略,包括在更小步长内调整λ_{REN},并结合学习率和权重衰减等超参数优化模型表现。

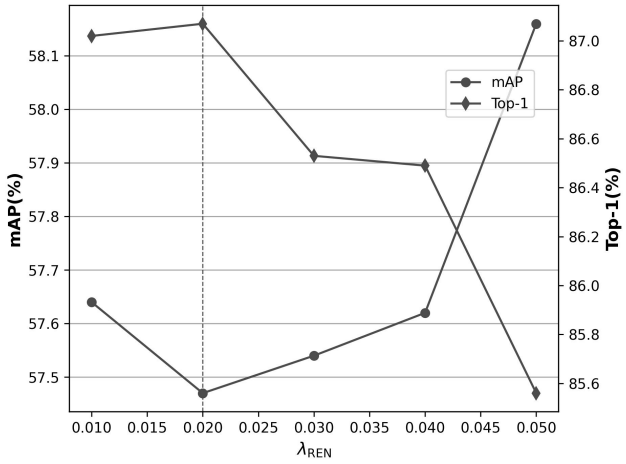

图4 平衡系数阈值实验分析

为探究增强头模块对模型架构的影响,本文在自身网络结构中的两个模块上进行了详细的补充实验。实验结果如表4所示,通过加入增强头模块,可以显著提升匹配的准确度。同时三阶段平均检测结果相比两阶段出现了降低,说明该端到端网络能够有效调整检测阶段和匹配阶段的冲突,使网络的重心偏向主任务匹配,从而实现更加精确的搜索效果。

表4 重识别头与增强头的性能比较(%)

方法	检测效果		搜索效果		
	recall	ap	mAP	top-1	top-10
重识别头	**95.12**	**90.96**	57.47	87.07	95.19
+增强头	93.41	88.22	**58.29**	**87.21**	**95.43**

2.5 对比实验

如表5所示,本文在CUHK-SYSU数据集和PRW数据集上比较了本文的SwinCascador方法和最先进的方法,包括两阶段方法[1-3]和端到端方法[17-21]等。

在CUHK-SYSU数据集上,本文方法取得了95.15%的mAP和95.79%的top-1准确率,相较于将检测和重识别分为两个细化模块的最佳Two-step方法TCTS[3]表现更优。在端到端方法中,本文方法相较于两级序列模型SeqNet[15]展现了显著的性能优势。在采用Transformer架构的相关方法中,本文方法的性能也优于最先进的SOIM[19]。值得注意的是,SOLIDER[21]作为一种以人为中心的通用模型,我们与其结合下游行人搜索任务模型SeqNet[15]的专用实验结果进行了对比,进一步验证了本文方法的有效性。

在PRW数据集上,由于数据集中存在显著的姿态/视角变化以及跨摄像机设置,该任务更具挑战性。但是,本文方法在mAP和top-1准确率上优于现有方法,其中mAP指标相较于最优方法提高了1.45%,展现了显著的性能优势。本文将这一表现归因于其逐级联的细化设计,通过逐步优化,为重识别任务提供了更高质量的特征表示。然而,本文方法在top-1搜索准确率上与表5中的PSTR[11]相比仍存在细微差距。这是因为在使用角度和范数信息进行特征区分时,本文更注重不同身份特征的差异化表达,而对同一身份内特征的一致性关注相对不足,从而对top-1准确率产生了一定影响。

表5 在CUHK-SYSU和PRW数据集中与其他方法比较(%)

方法		PRW		CUHK-SYSU	
		mAP	top-1	mAP	top-1
两阶段	IGPN[1]	47.2	87.0	90.3	91.4
	RDLR[2]	42.9	70.2	93.0	94.2
	TCTS[3]	46.8	87.5	93.9	95.1
端到端	基于CNN的方法				
	OIM[5]	21.3	49.9	75.5	78.7
	NAE[17]	43.3	80.9	91.5	92.4
	AlignPS[7]	45.9	81.9	93.1	93.4
	SeqNet[15]	46.7	83.4	93.8	94.6
	VSRI[18]	52.9	87.3	93.4	94.1
	PAPS[8]	51.6	74.7	92.1	94.0
	基于Transformer的方法				
	PSTR[11]	49.5	**87.8**	93.5	95.0
	COAT[12]	53.3	87.4	94.2	94.7
	SOIM[19]	52.4	85.4	93.6	94.2
	COAT+HKD[20]	53.49	86.63	93.86	94.76
	SOLIDER[21]	56.84	86.78	94.91	95.72
	本文	**58.29**	87.21	**95.15**	**95.79**

表6 使用后处理操作(上下文二部图匹配)(%)

方法	mAP	top-1
AlignPS[7]+CBGM[15]	46.8	85.8
SeqNet+CBGM[15]	47.6	87.6
COAT[12]+CBGM[15]	54.0	**89.1**
本文+CBGM[15]	**58.91**	88.82

本文使用后处理操作利用匹配算法对候选结果进行重新排序,显著提高了top-1的准确率,在表6中显示了使用上下文二部图匹配(CBGM[15])后处理的操作,本文的方法在map精度方面比所有现有方法都要好。如表7所示,本文的性能优于其他使用Swin Transformer tiny作为骨干网络的方法。图5展示了在同一行人出现尺度变化较大时仍能准确的搜索到目标人物。

表 7 使用 Swin Transformer tiny 作为主干(%)

方法	PRW		CUHK-SYSU	
	mAP	top-1	mAP	top-1
SOIM[19]	52.4	85.4	93.6	94.2
SOLIDER[21]	56.84	86.78	94.91	95.72
本文	**58.29**	**87.21**	**95.15**	**95.79**

3 结束语

本文提出了一种基于 Swin Transformer 的端到端级联框架,称之为 SwinCascador。通过使用多阶段级联作为一种重采样机制,本文实现了每个阶段专注于不同任务的能力,从而减轻两个任务之间的潜在冲突。本文还增加了额外的监督项(AEN),以抑制推理阶段无法对这些提议框进行采样的影响,降低同一身份下特征向量的差异。通过广泛的消融研究,本文验证了每个组件的有效性,并展示了在行人搜索基准上相较于其他方法的显著优势。

为进一步提升 SwinCascador 框架的整体性能,未来工作将着重优化阶段间的特征一致性与信息传递效率。在各阶段增加特征对齐模块,以统一空间与语义特征,同时通过权重共享减少冗余,提升协同优化能力。此外,引入反馈机制,使后续阶段的优化结果反作用于前一阶段,形成闭环的优化过程,以提升框架的整体表现。

(a)搜索目标行人 (b) top-1 (c) top-2 (d) top-3

图 5 实验结果可视化

参 考 文 献

[1] Dong W, Zhang Z, Song C, et al. Instance guided proposal network for person search[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 2585-2594.

[2] Han C, Ye J, Zhong Y, et al. Re-id driven localization refinement for person search[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 9814-9823.

[3] Wang C, Ma B, Chang H, et al. Tcts: A task-consistent two-stage framework for person search[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 11952-11961.

[4] 柯健宇,王晓峰.注意力机制和多元损失改进的行人重识别模型[J].计算机应用与软件,2024,41(03):174-181.

[5] Xiao T, Li S, Wang B, et al. Joint detection and identification feature learning for person search[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 3415-3424.

[6] Li Z, Miao D. Sequential end-to-end network for efficient person search[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2021, 35(3): 2011-2019.

[7] Yan Y, Li J, Qin J, et al. Efficient person search: An anchor-free approach[J]. International Journal of Computer Vision, 2023, 131(7): 1642-1661.

[8] Gaikwad B, Karmakar A. Real-time distributed video analytics for privacy-aware person search[J]. Computer Vision and Image Understanding, 2023, 234: 103749.

[9] Zhang P, Yu X, Bai X, et al. Joint discriminative representation learning for end-to-end person search[J]. Pattern Recognition, 2024, 147: 110053.

[10] Dosovitskiy A. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arxiv preprint arxiv:2010.11929, 2020.

[11] Cao J, Pang Y, Anwer R M, et al. Pstr: End-to-end one-step person search with transformers[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 9458-9467.

[12] Yu R, Du D, LaLonde R, et al. Cascade transformers for end-to-end person search[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 7267-7276.

[13] 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. 2021: 10012-10022.

[14] Chen D, Zhang S, Ouyang W, et al. Person search via a mask-guided two-stream cnn model[C]//Proceedings of the european conference on computer vision. 2018: 734-750.

[15] Li Z, Miao D. Sequential end-to-end network for efficient person search[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2021, 35(3): 2011-2019.

[16] Yu R, Du D, LaLonde R, et al. Cascade transformers for end-to-end person search[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 7267-7276.

[17] Chen D, Zhang S, Yang J, et al. Norm-aware embedding for efficient person search[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 12615-12624.

[18] Liu Y, Li Y, Kong C, et al. Vision Shared and Representation Isolated Network for Person Search[C]//IJCAI. 2022: 1216-1222.

[19] Xiang X, Lv N, Qiao Y. Transformer-based person search model with symmetric online instance matching[C]//ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022: 2729-2733.

[20] Zhang S, Yang Q, Cheng D, et al. Ground-to-aerial person search: Benchmark dataset and approach[C]//Proceedings of the 31st ACM International Conference on Multimedia. 2023: 789-799.

[21] Chen W, Xu X, Jia J, et al. Beyond appearance: a semantic controllable self-supervised learning framework for human-centric visual tasks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 15050-15061.