基于隐式解码对齐的空地行人重识别方法

贝俊仁 [1] 张 权 [1] 赖剑煌 [1, 2, 3, 4]

摘 要 空地行人重识别任务旨在包含地面与空中视角的监控相机网络中, 实现对特定行人的精确识别与跨镜关联. 该任务的特有挑战在于克服空地成像设备之间巨大的视角差异对于学习判别性行人身份特征的干扰. 现有工作在行人特征建模方面存在不足, 未充分考虑跨视角特征对齐对识别与检索性能的提升作用. 基于此, 提出一种基于隐式解码对齐的空地行人重识别方法, 主要包含两方面的创新: 在模型设计方面, 提出基于自注意力解码器的隐式对齐框架, 通过在解码阶段利用一组可学习的口令特征挖掘行人判别部件区域, 并提取和对齐行人局部特征, 从而实现判别性行人表征的学习; 在优化目标方面, 提出正交性和一致性损失函数, 前者约束口令特征以多样化判别性行人部件为关注点, 后者缓解了跨视角特征表达的偏置分布. 在当前可用的最大空地重识别数据集 CARGO 上进行实验, 结果表明所提方法在检索性能上优于现有重识别方法, 实现显著的性能提升.

关键词 行人重识别, 图像检索, 图像识别, 自注意力网络

引用格式 贝俊仁, 张权, 赖剑煌. 基于隐式解码对齐的空地行人重识别方法. 自动化学报, 2025, **51**(9): 1988−2000

DOI 10.16383/j.aas.c240705 **CSTR** 32138.14.j.aas.c240705

Implicit Decoder Alignment for Aerial-ground Person Re-identification

BEI Jun-Ren[1] ZHANG Quan[1] LAI Jian-Huang[1, 2, 3, 4]

Abstract Aerial-ground person re-identification (AGPReID) refers to the accurate recognition and cross-camera association of the interest person within a surveillance camera network deployed across aerial and ground views. One of the unique challenges is overcoming the significant view differences between aerial and ground cameras, which interfere with discriminate feature learning of person identity. Existing methods are insufficient in modeling person features, as they fail to fully leverage cross-view feature alignment to enhance recognition and retrieval performance. To this end, the implicit decoder alignment is proposed for AGPReID with two main contributions: In terms of model design, we propose an implicit alignment framework based on a self-attention decoder. This framework leverages a set of learnable token features during the decoding stage to identify discriminative person parts and to extract and align local pedestrian features, enabling discriminative person representation learning. In terms of optimization objectives, we propose orthogonality and consistency loss functions. The former constrains token features to focus on diverse discriminative pedestrian parts, while the latter mitigates biased distributions in cross-view feature representations. Experiments were conducted on the largest publicly available cross-view re-identification dataset, CARGO, where the proposed method outperformed existing re-identification methods, achieving superior retrieval performance.

Key words Person re-identification, image retrieval, image recognition, transformer

Citation Bei Jun-Ren, Zhang Quan, Lai Jian-Huang. Implicit decoder alignment for aerial-ground person re-identification. *Acta Automatica Sinica*, 2025, **51**(9): 1988−2000

行人重识别 (Person re-identification, ReID) 任务旨在非重叠视域的多摄像头监控网络中实现对特定对象的准确识别和跨摄像头跟踪[1−7]. 随着“智慧城市”和“雪亮工程”等项目的实施, 在深度学习和计算机视觉的支持下, 该技术已在城市安防、边境监控、园区管理等领域中发挥关键作用. 尽管现有研究在纯地面视角[8−12] 和纯无人机视角[13−17] 的监控网络中取得显著进展, 但这些研究往往假设监控

收稿日期 2024-10-31 录用日期 2025-05-29
Manuscript received October 31, 2024; accepted May 29, 2025
国家自然科学基金 (U22A2095), 国家资助博士后研究人员计划中国博士后科学基金 (GZC20252314), 广州市重点研发计划 (202206030003), 广东省信息安全技术重点实验室项目 (2023B1212060026) 资助
Supported by National Natural Science Foundation of China (U22A2095), the Postdoctoral Fellowship Program and China Postdoctoral Science Foundation (GZC20252314), Key-Area Research and Development Program of Guangzhou (202206030003), and the Project of Guangdong Provincial Key Laboratory of Information Security Technology (2023B1212060026)
本文责任编委 鲁继文
Recommended by Associate Editor LU Ji-Wen
1. 中山大学 广州 510275 2. 琶洲实验室 (黄埔) 广州 510006 3. 广东省信息安全技术重点实验室 广州 510006 4. 机器智能与先进计算教育部重点实验室 广州 510006
1. Sun Yat-sen University, Guangzhou 510275 2. Pazhou Laboratory (Huangpu), Guangzhou 510006 3. Guangdong Province Key Laboratory of Information Security Technology, Guangzhou 510006 4. Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, Guangzhou 510006

系统由相同类型的摄像头组成, 成像风格较为一致, 因此可以将问题简化为同构相机网络下的图像检索任务. 然而, 这一简化与现实中的监控网络存在显著差距, 实际部署的监控系统通常包含不同类型和视角的摄像头.

为解决上述不足, 本文关注一种更贴近实际应用且更具挑战性的空地行人重识别任务 (Aerial-ground person re-identification, AGPReID), 即在监控网络中同时融合地面摄像头和无人机摄像头的视角. 该任务不仅拓宽行人重识别的研究范围, 也更加符合实际监控的需求. 如图 1 所示, 地面摄像头覆盖人流密集的城市区域, 而无人机摄像头则覆盖市郊、农田、森林等人流稀疏的区域. 这种组合不仅提升了传统行人重识别研究的适用性, 还显著扩展了其在实际部署场景中的应用潜力. 空地行人重识别任务作为现有研究的重要补充和扩展, 提供新的研究视角, 助力实现更全面的监控覆盖和更高效的目标识别检索. 与传统行人重识别相比, 该任务的独特挑战在于跨镜头视角差异显著, 这对判别性身份表征的学习带来较大干扰. 然而, 空地行人重识别研究仍处于起步阶段, 现有方法未能有效解决这一挑战. 此外, 直接迁移传统行人重识别方法[18−20]至该场景会导致明显的性能下降, 因为空地视角差异引起的行人形变使得传统的对齐方法难以生效. 这一现状促使设计新的对齐方法, 以缓解空地视角间的巨大差异.

针对上述问题, 本文提出一种基于隐式特征对齐的空地行人重识别方法, 主要在模型设计和优化目标上进行创新: 1) 在模型设计上, 提出基于解码对齐的纯自注意力框架 IDA (Implicit decoder alignment framework), 该框架包含自注意力编码器和解码器模块, 编码器负责提取输入图像的特征, 解码器通过设计一组可学习的部件对齐口令特征, 对编码器特征进行查询、挖掘和对齐, 从而识别出关键行人部件区域, 实现判别性行人表征的学习; 2) 在优化目标上, 设计正交性损失和一致性损失, 用于监督口令特征的学习, 其中正交性损失约束口令特征关注不同且具有判别性的行人部件区域, 一致性损失调整视角间的特征相似度分布与视角内的特征相似度分布一致, 从而缓解跨视角特征的有偏分布.

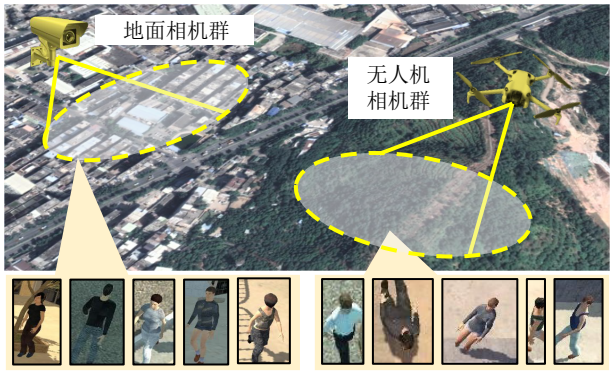


图 1　空地行人重识别任务示意图

Fig. 1　Illustration of the aerial-ground person re-identification task

最后, 在当前最大规模的空地行人重识别数据集 CARGO 上进行实验验证. 结果表明, 所提出的 IDA 方法在多种空地检索模式下均表现优异, 尤其是在跨空地重识别设定中, mINP 实现 5.04% 的提升, 显著超越当前最佳方法.

1　相关工作概述

1.1　传统行人重识别

传统行人重识别主要是指基于同构视角的行人重识别任务, 进一步分为地面视角下和无人机视角下的行人重识别. 随着深度学习的日益发展, 传统基于手工特征和专家先验[21−24]的方法表征能力受限, 基于深度学习的表征学习[25−35]和度量学习已经成为主流的研究策略[36−47].

在地面视角的相机系统中, Ye 等[18]在上述基础上进行改进, 提出基于卷积神经网络的强力模型 AGW (Attention generalized mean pooling with weighted triplet loss), 其中设计一个新的带权正则三元组损失函数, 从而避免引入和调整额外超参数. He 等[20]在其基础上进一步改进, 加入非局部注意力模块 (Non-local attention block) 和可泛化均值池化 (Generalized mean pooling, GeM), 进一步提升模型的性能. Zhou 等[27−28]在模型构建的基础上考虑轻量化设计, 提出全尺度特征学习网络 (Omni-scale feature learning network, OSNet), 其在卷积网络中设计多尺度特征的提取与聚合, 保持特征判别性的同时大幅度减少模型的参数量. Luo 等[33]提出一种简单有效的基于卷积神经网络的基线模型, 其中在特征提取网络的最后一层缩小步长从而扩大感受野. 除此之外, 也有一系列的工作注意到细粒度的局部特征对于行人身份判别的重要性. Sun 等[25−26]提出基于特征硬分块的行人重识别模型重识别模型 PCB (Part-based convolutional baseline), 其对卷积网络提取得到的行人特征进行

从上到下的特征水平分块操作, 并约束模型在每一块中捕捉与身份相关的细粒度特征. Luo 等[34] 在此基础上设计分块之间的对齐策略, 利用动态规划思想解决行人尺度不一致导致的特征不匹配问题. Wang 等[31] 提出多粒度网络 (Multiple granularity network, MGN), 其设计更多尺度的分块策略用于关注判别性细粒度行人特征. Zhang 等[12] 设计一种基于软分块策略的行人全局与局部特征提取方法, 避免已有硬分块策略中特征难以匹配对齐的不足. 在自注意力模型兴起之后, He 等[35] 也提出第一个基于视觉自注意力模型的行人重识别网络 (Transformer-based object re-identification, TransReID), 其中引入摄像头信息进一步增强模型的判别能力. 然而, 上述方法仅考虑纯地面视角下的相机网络系统, 对于更复杂场景下的视角变化与行人特征对齐是无法良好解决的.

面向基于无人机监控相机系统的行人重识别研究[14−15, 48−56] 是相对较少的. 其中, Wang 等[14] 提出一种基于旋转探索的自注意力框架 (Rotation exploration transformer, RoExViT), 用于解决无人机视角下前景行人的旋转和背景风格的变化问题. Qiu 等[15] 提出一种基于前景显著性特征和行人关键点信息在无人机视角下对齐行人特征的方法 SPAKD (Salient part-aligned and keypoint disentangling framework). Zhang 等[48] 提出基于子空间池化的方法, 用于提取无人机视角下最具判别性的前 k 个行人特征. 然而, 上述方法仅关注无人机视角下行人的表观变化, 对于空地更复杂场景下的行人变化考虑不充分, 因此直接迁移无法直接获得良好的识别性能.

1.2 空地行人重识别

与上述传统任务不同, 空地行人重识别考虑同时部署在地面和无人机视角下的监控相机系统, 属于异构视角下的行人图像检索问题. 该任务面临的挑战在于不仅要解决每种相机下的行人特征提取问题, 还需要考虑行人表现在经历显著跨视角后的特征对齐问题.

目前该任务的研究仍处于起步阶段. Nguyen 等[50] 是第一个从深度学习的视角考虑该任务的先驱性工作, 在标注基于自注意力的行人重识别方法上额外引入一个可解释性重识别分支, 基于属性标签对于成对行人图像的相似性判断给出可靠的解释. 除此之外还贡献一个空地重识别数据集, 包含 388 个行人身份和 21 893 个行人图像, 但是该数据集的采集规模太小, 无法充分发挥深度模型的训练优势. Zhang 等[51] 提出一个基于自注意力模型的视角解耦框架 (View-decoupled transformer, VDT), 其通过在自注意力编码器中额外引入视角分类任务, 从而使得模型解耦视角相关与视角无关的特征, 并缓解视角差异对于行人身份特征的干扰. 除此之外, 他们也贡献了当前最大规模的空地重识别数据集 CARGO, 其身份规模和数据规模分别是已有数据集的约 10 倍和 5 倍. 然而, 上述的研究与本文方法在研究路线、网络结构与优化策略方面存在突出差异. 1) 研究路线. 针对空地相机网络场景下存在的显著视角差异挑战, 本文与 VDT 采取截然不同的研究路线. VDT 方法主要从特征解耦的角度出发, 旨在获得与视角无关的身份特征表达. 然而, 这种方法对行人细粒度特征的表达与对齐问题关注不足. 相比之下, 本文从特征对齐的角度切入, 重点学习和挖掘空地跨视角场景下共有的多样化行人部件区域特征, 从而实现更精确的识别与检索性能. 2) 网络结构. 为支撑各自的研究路线, 本文与 VDT 在网络架构设计上存在本质区别. VDT 采用一个基于 Transformer 自注意力模型的纯编码器架构. 而本文方法构建一个包含编码器−解码器的完整 Transformer 架构, 其中编码器负责特征提取, 解码器则专注于特征对齐, 这种双模块设计能够更有效地实现跨视角特征的隐式对齐. 3) 优化策略. VDT 使用正交损失函数, 通过计算余弦相似度来约束身份特征与视角特征之间的独立性. 而本文方法的正交性损失通过归一化特征矩阵的 F 范数来约束多个局部特征之间的两两独立性, 这种约束形式更具泛化性和通用性; 其次, 特别设计一个一致性损失函数, 用于约束跨视角特征表达的有偏分布, 从而进一步提升模型在复杂场景下的鲁棒性.

1.3 模型架构

在当前的重识别研究中, 基于深度学习的模型框架是主流思路, 并且其中大部分工作都是以卷积神经网络[29] 为主. 随着自注意力模型[52] 在自然语言任务中大获成功以来, Vaswani 等[53] 将其完善为视觉自注意力模型 (Vision transformer, ViT), 并在多个视觉任务上获得更优越的性能. 基于此, 已有的许多行人重识别方法也基于 ViT 展开. 然而上述工作在进行表征学习时对于自注意力中的编码阶段进行充分的设计, 对于解码阶段的探究是不足够的. Dosovitskiy 等[54] 的工作也指出, 在编码器阶段学习局部特征时是存在注意力偏置的. 基于此, 本文将编码器和解码器纳入一个完整的框架中, 并且在编码器阶段实现全局特征的提取, 在解码器阶段实

现局部特征的提取与对齐, 从而获得空地场景下的判别性行人表征.

2　方法

2.1　任务建模

给定一个空地视角下的行人重识别数据集 $\mathcal{D}=\{(x_i,\ y_i,\ v_i)\}_{i=1}^{|\mathcal{D}|}$, 其由训练集 $\mathcal{D}^{tr}$ 和测试集 $\mathcal{D}^{te}$ 两部分组成, 其中 x_i 代表第 i 个行人图像, y_i 代表对应的行人身份标签, v_i 代表行人图像的视角标签 (0-1 二值标签), 即该图像来自地面视角还是无人机视角. 视角标签是不需要额外标注的, 可以根据采集行人图像时的摄像头编号轻易获得.

空地行人重识别任务的目标在于构建一个深度模型 $\mathcal{F}(\cdot;\theta)$, 使其对于任一输入的行人图像 x_i, 都能学习到具有判别力的身份表征, 最终使得特征空间呈现良好的身份内聚性, 可描述为如下形式:

$$\min_{\theta}\sum_{x_i\in\mathcal{D}}\left[\left\|\mathcal{F}\left(x_i;\theta\right)-\mathcal{F}\left(x_i^{+};\theta\right)\right\|_2-\left\|\mathcal{F}\left(x_i;\theta\right)-\mathcal{F}\left(x_i^{-};\theta\right)\right\|_2\right] \tag{1}$$

其中, x_i^{+} 和 x_i^{-} 分别代表 x_i 的同身份正样本和不同身份负样本, θ 为对应模型 $\mathcal{F}$ 中的可学习参数. 而由于不同视角标签之间的显著差异, 导致稳定鲁棒的行人特征是难以提取和对齐的. 因此在本工作中, 从模型设计 $\mathcal{F}$ 出发, 将其拆解为基于自注意力模型的编码器 $\mathcal{F}_e(\cdot;t_g,\ \theta_e)$ 和解码器 $\mathcal{F}_d(\cdot;t_a^1,\ \cdots,\ t_a^{N_q},\ \theta_d)$, 通过引入一系列的可学习口令特征, 实现局部特征的挖掘与对齐, 从而更好地实现式 (1). 本文中常用的符号及其释义总结在表 1 中.

2.2　模型结构

本文所提出的方法结构如图 2 所示, 主要包含输入图像采样、基于自注意力的隐式解码对齐框架 (涉及特征提取以及特征对齐)、以及损失函数监督 (涉及特征优化). 本节主要介绍输入采样以及框架结构, 对于损失函数的介绍详见第 2.3 节.

表 1　常用符号及其含义

Table 1　Commonly used symbols and their meanings in our work

符号	含义
$\mathcal{D},\ \mathcal{D}^{tr},\ \mathcal{D}^{te}$	空地重识别数据集, 训练集, 测试集
B	批数据
$x_i,\ y_i,\ v_i$	行人图像, 身份标签, 视角标签
$\mathcal{F},\ \mathcal{F}_e,\ \mathcal{F}_d$	网络模型, 自注意力编码器和解码器
$\boldsymbol{Q},\ \boldsymbol{K},\ \boldsymbol{V}$	注意力操作的查询、键值和内容
$\theta,\ \theta_e,\ \theta_d$	$\mathcal{F},\ \mathcal{F}_e,\ \mathcal{F}_d$ 中的可学习参数
$\mathcal{T}(\cdot)$	输入离散口令化
$t_g,\ t_a$	可学习口令特征
$\boldsymbol{L}$	局部口令特征矩阵
$\boldsymbol{S}_{g\leftrightarrow g},\ \boldsymbol{S}_{a\leftrightarrow a},\ \boldsymbol{S}_{a\leftrightarrow g}$	相似度矩阵
$\mathcal{L}_g^c,\ \mathcal{L}_g^t$	全局特征损失函数
$\mathcal{L}_a^c,\ \mathcal{L}_a^t,\ \mathcal{L}_a^o,\ \mathcal{L}_a^s$	局部特征损失函数
$\lvert\cdot\rvert$	集合的阶

输入采样. 对于模型训练输入的批处理数据, 设计跨视角平衡采样的策略, 即在均匀采样身份的基础上, 进一步均匀采样同一身份来自地面视角和无人机视角的行人图像. 具体地, 对于批处理数据 B, 每次从训练集 $\mathcal{D}^{tr}$ 中随机采样 P 个行人身份, 对于每个行人身份, 随机采样 K 个行人图像, 其中确保 $K/2$ 个图像来自地面视角, 另外 $K/2$ 个图像来自无人机视角. 因此, 一次训练的批数据大小为 $|B|=P\times K$.

隐式解码对齐框架 IDA. 该框架 IDA 由自注意力编码模块 $\mathcal{F}_e(\cdot;t_g,\ \theta_e)$ 和自注意力解码模块

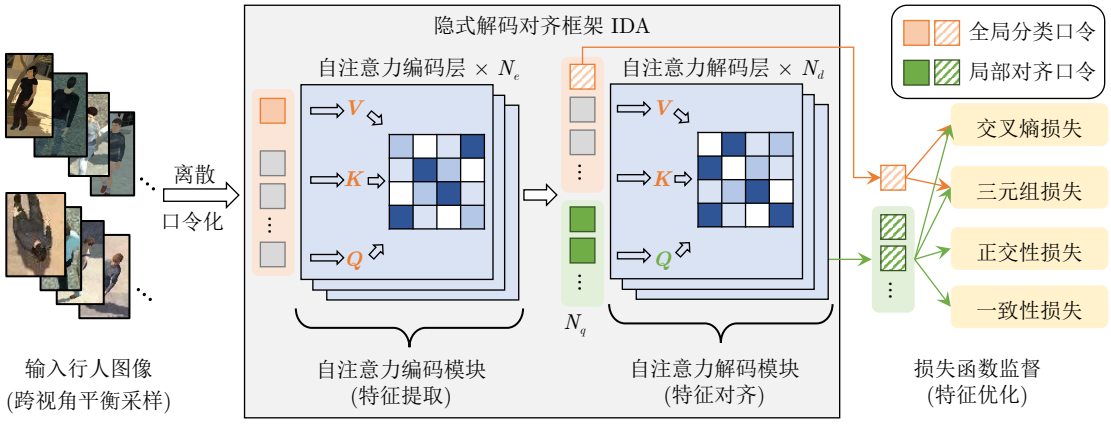


图 2　隐式解码对齐框架示意图

Fig. 2　Illustration of the implicit decoder alignment framework

$\mathcal{F}_d(\cdot; t_a^1, \cdots, t_a^{N_q}, \theta_d)$ 两个部分组成, 可记为:

$$\mathcal{F}(x_i) = \mathcal{F}_d\left(\mathcal{F}_e(x_i)\right) \tag{2}$$

其中, t_g 与 t_a 分别代表模块内部可学习的口令特征, θ_e 与 θ_d 代表编码器和解码器的全部可学习参数. $\mathcal{F}_e$ 和 $\mathcal{F}_d$ 分别由 N_e 个自注意力编码层和 N_d 个解码层堆叠组成, 每一层的结构与经典自注意力模型[52] 保持一致, 每层中的核心操作均为注意力机制 $\mathcal{A}(\cdot)$, 可描述为:

$$\mathcal{A}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \delta\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{3}$$

其中, $\delta(\cdot)$ 为 SoftMax 函数, $\sqrt{d_k}$ 为放缩因子常数. 除此之外, 在 $\mathcal{F}_e$ 引入一个可学习的全局分类口令特征 t_g, 用于提取输入图像的全局身份表征; 在 $\mathcal{F}_d$ 引入 N_q 个可学习的局部对齐口令特征 t_a^i, $1 \le i \le N_q$, 用于在编码层特征的基础上, 进一步挖掘具有判别性的行人局部区域特征. 最后, 将全局口令与局部口令拼接, 作为输入图像的完整的特征表达.

接下来是对 IDA 具体流程的介绍. 对于任意 $x_i \in B$, 首先将其离散化, 然后通过 1×1 的卷积层投影为嵌入特征, 接着与编码器中的全局口令拼接在一起经过位置编码, 得到 IDA 的编码输入. 可以将上述过程定义为:

$$\mathcal{T}(x_i) := [t_g(x_i), e_{x_i}^1, \cdots, e_{x_i}^j, \cdots, e_{x_i}^M] \tag{4}$$

其中, $\mathcal{T}$ 代表对 x_i 的离散口令化, $\mathcal{T}(x_i) \in \mathbf{R}^{(M+1)\times d}$, $e_{x_i}^j$ 代表由离散图像块变换得到的 d 维嵌入特征. 在每一个编码层中, 对于输入序列执行自注意力操作. 以第一层为例, 即对于式 (3), 其输入分别是:

$$\begin{cases} \boldsymbol{Q} = \mathcal{T}(x_i)\boldsymbol{W}_Q^{\mathcal{F}_e^1} \\ \boldsymbol{K} = \mathcal{T}(x_i)\boldsymbol{W}_K^{\mathcal{F}_e^1} \\ \boldsymbol{V} = \mathcal{T}(x_i)\boldsymbol{W}_V^{\mathcal{F}_e^1} \end{cases} \tag{5}$$

其中, $\boldsymbol{W}_Q^{\mathcal{F}_e^1}$, $\boldsymbol{W}_K^{\mathcal{F}_e^1}$, $\boldsymbol{W}_V^{\mathcal{F}_e^1}$ 是编码层第一层中的三个不共享的自注意力变换矩阵. 在经过 N_e 个编码层操作之后, IDA 会得到 x_i 的全局编码表征, 可记为:

$$\mathcal{F}_e(\mathcal{T}(x_i)) = [t_g^{N_e}(x_i), e_{x_i}^{1,N_e}, \cdots, e_{x_i}^{M,N_e}] \tag{6}$$

接下来, 将编码器得到的特征送入解码器 $\mathcal{F}_d$ 中, 进一步挖掘和对齐输入图像的行人局部特征. 在解码器中引入一系列可学习的局部对齐口令特征, 可记为 $\boldsymbol{L} = [t_a^1, \cdots, t_a^{N_q}] \in \mathbf{R}^{N_q \times d}$. 在解码器的每一层中, 最核心的操作为跨注意力操作. 以第一层解码层为例, 对于式 (3), 其输入分别是:

$$\begin{cases} \boldsymbol{Q} = \boldsymbol{L}\boldsymbol{W}_Q^{\mathcal{F}_d^1} \\ \boldsymbol{K} = \mathcal{F}_e(\mathcal{T}(x_i))\boldsymbol{W}_K^{\mathcal{F}_d^1} \\ \boldsymbol{V} = \mathcal{F}_e(\mathcal{T}(x_i))\boldsymbol{W}_V^{\mathcal{F}_d^1} \end{cases} \tag{7}$$

其中, $\boldsymbol{W}_Q^{\mathcal{F}_d^1}$, $\boldsymbol{W}_K^{\mathcal{F}_d^1}$, $\boldsymbol{W}_V^{\mathcal{F}_d^1}$ 是解码层第一层中的三个不共享的自注意力变换矩阵. 在经过 N_d 层解码操作之后, 局部对齐口令通过与编码特征不断交互, 在损失函数的监督作用下最终收敛于多个具有判别性的行人局部区域, 可记为:

$$\mathcal{F}(x_i) = \mathcal{F}_d\left(\mathcal{F}_e(x_i)\right) = \left[t_a^{1,N_d}(x_i), \cdots, t_a^{j,N_d}(x_i) \cdots, t_a^{N_p,N_d}(x_i)\right] \tag{8}$$

其中, $t_a^{j,N_d}(x_i)$, $1 \le j \le N_d$ 代表 t_a^j 在经过解码模块的最后一层后关于 x_i 的局部口令特征, $\mathcal{F}(x_i) \in \mathbf{R}^{N_q \times d}$. 最终, 将全局分类口令特征 $t_g^{N_e}(x_i)$ 和局部对齐口令特征 $\mathcal{F}(x_i)$ 用于损失函数的监督训练. 在测试阶段, 将上述两种特征拼接在一起作为完整的行人特征表达.

2.3 模型优化

完整的 IDA 优化流程展示在算法 1 中. 在得到模型输出的两种口令特征之后, 对两种口令特征分别施加损失函数进行监督训练. 对于全局口令特征 $t_g^{N_e}(x_i)$, 分别使用交叉熵和三元组损失函数进行监督, 其形式具体为:

$$\mathcal{L}_g^c = -\frac{1}{|B|}\sum_{i=1}^{|B|} y_i \log_2 p\left(y_i \mid t_g^{N_e}(x_i)\right) \tag{9}$$

其中, y_i 是 x_i 的真实身份标签, $p\left(y_i \mid t_g^{N_e}(x_i)\right)$ 代表基于 x_i 的全局口令特征进行身份分类的预测值.

$$\mathcal{L}_g^t = \frac{1}{|B|}\sum_{i=1}^{|B|}\Big[\left\|t_g^{N_e}(x_i) - t_g^{N_e}\left(x_i^+\right)\right\|_2^2 - \left\|t_g^{N_e}(x_i) - t_g^{N_e}\left(x_i^-\right)\right\|_2^2 + \alpha\Big] \tag{10}$$

其中, x_i^+ 和 x_i^- 代表当前样本 x_i 的困难正样本和困难负样本, α 代表调节正负样本对之间距离的超参数.

算法 1. IDA 训练流程算法

输入. 训练集 $\mathcal{D}^{tr}$, 模型编码器 $\mathcal{F}_e(\cdot; t_g, \theta_e)$ 和解码器 $\mathcal{F}_d(\cdot; t_a^1, \cdots, t_a^{N_q}, \theta_d)$

输出. 最优网络参数 $\theta_e^\star$ 和 $\theta_d^\star$

初始化 $\theta_{\mathcal{N}}$ 和 $\theta_{\mathcal{F}}$;

while $\mathcal{F}_e$ 和 $\mathcal{F}_d$ 不收敛 **do**

  跨视角平衡采样 $B = \{(x_i, y_i, v_i)\}_{i=1}^{|B|}$ 从 $\mathcal{D}^{tr}$ 中;

  for $x_i \in B$ **do**

5　　提取 $\mathcal{T}(x_i)$ 通过式 (4);
6　　提取全局口令在 $\mathcal{F}_e$ 通过式 (6);
7　　提取局部口令在 $\mathcal{F}_d$ 通过式 (7);
8　　**end for**
9　计算 $\mathcal{L}_g^c$, $\mathcal{L}_g^t$ 通过式 (9) 和式 (10);
10　计算 $\mathcal{L}_a^c$, $\mathcal{L}_a^t$ 通过式 (9) 和式 (10);
11　计算 $\mathcal{L}_a^o$ 通过式 (11);
12　计算 $\mathcal{L}_a^s$ 通过式 (12), 式 (13) 和式 (14);
13　计算 $\mathcal{L}$ 通过式 (15);
14　**end while**
15 return 最优参数 $\theta_e^\star$ 和 $\theta_d^\star$

对于局部口令特征 $[t_a^{1,N_d}(x_i), \cdots, t_a^{N_p,N_d}(x_i)]$, 其每一项首先也要经过交叉熵和三元组损失的身份监督学习, 形式与式 (9) 和式 (10) 一致, 可记为 $\mathcal{L}_a^c$ 和 $\mathcal{L}_a^t$. 除此之外, 还设计正交化损失函数 $\mathcal{L}_a^o$ 和一致性损失函数 $\mathcal{L}_a^s$ 来进一步约束局部特征的分布. 其中, $\mathcal{L}_a^o$ 约束提取得到的局部特征在保持判别力的同时, 学习到彼此无关更多样化的特征表达, 可记为:

$$\mathcal{L}_a^o = \sum_{x_i \in B} \left\| \mathcal{F}^\star(x_i)\, \mathcal{F}^{\star\mathrm{T}}(x_i) - \boldsymbol{I} \right\|_{\mathrm{F}}^2 \tag{11}$$

其中, $\mathcal{F}^\star(x_i)$ 是逐行归一化后的局部特征矩阵, $\boldsymbol{I}$ 代表形状为 $N_q \times N_q$ 的单位对角矩阵, $\|\cdot\|_{\mathrm{F}}$ 代表矩阵的 F 范数. 值得注意的是, 与 VDT 中的正交损失函数[51] 相比, 本文所提出的 $\mathcal{L}_a^o$ 在实现策略和目标上仍存在以下几个关键差异: 1) 实现策略的不同. VDT 中主要关注两个特征之间的无关性表达, 通过约束特征间的余弦相似度来实现正交性约束. 而本文提出的 $\mathcal{L}_a^o$ 则关注多个特征之间的无关性表达, 通过约束归一化后的特征矩阵的 F 范数, 进而同时实现多个特征两两之间的正交性约束. 因此, 本文的损失函数提供一种更加泛化和通用的约束形式, 能够有效处理更复杂的特征关系. 2) 目标的不同. VDT 中的目标是约束身份特征与视角特征之间的无关性, 从而实现良好的特征解耦. 而本文提出的 $\mathcal{L}_a^o$ 的目标则是约束多个局部口令特征之间的无关性, 进而在空地视角下关注不同的行人区域, 确保特征对齐的精确性和一致性.

$\mathcal{L}_a^s$ 进一步约束局部特征在视角内的相似度分布与视角间的相似度分布趋于一致, 从而缓解视角差异. 具体来说, 首先将批数据 B 特征按照视角标签进行分类, 可表示为:

$$\begin{cases} \boldsymbol{E}_g = \{\mathcal{F}^\star(x_i) \mid x_i \in B \wedge v_i = 0\} \\ \boldsymbol{E}_a = \{\mathcal{F}^\star(x_i) \mid x_i \in B \wedge v_i = 1\} \end{cases} \tag{12}$$

值得注意的是, 尽管在式 (12) 中采用集合形式挑选特征, 但是最后会以矩阵的形式组织成为 $\boldsymbol{E}_a$ 和 $\boldsymbol{E}_g$. 对于两类数据, 分别有三种相似度关系矩阵, 记为如下形式:

$$\boldsymbol{S}_{g\leftrightarrow g} = \boldsymbol{E}_g \boldsymbol{E}_g^{\mathrm{T}},\ \boldsymbol{S}_{a\leftrightarrow a} = \boldsymbol{E}_a \boldsymbol{E}_a^{\mathrm{T}},\ \boldsymbol{S}_{a\leftrightarrow g} = \boldsymbol{E}_a \boldsymbol{E}_g^{\mathrm{T}} \tag{13}$$

其中, $\boldsymbol{S}_{g\leftrightarrow g}$ 和 $\boldsymbol{S}_{a\leftrightarrow a}$ 计算视角内相似度, $\boldsymbol{S}_{a\leftrightarrow g}$ 计算视角间相似度. 为了实现相似度约束, 以高斯分布建模 $\mathrm{N}(\mu, \sigma^2)$ 相似度关系, 并约束彼此的均值和标准差接近, 实现分布一致性拉近, 可记为:

$$\begin{aligned} \mathcal{L}_a^s = &\left(\|\mu_{g\leftrightarrow g} - \mu_{a\leftrightarrow a}\|_2^2 + \|\sigma_{g\leftrightarrow g} - \sigma_{a\leftrightarrow a}\|_2^2 \right) + \\ &\left(\|\mu_{g\leftrightarrow g} - \mu_{a\leftrightarrow g}\|_2^2 + \|\sigma_{g\leftrightarrow g} - \sigma_{a\leftrightarrow g}\|_2^2 \right) + \\ &\left(\|\mu_{a\leftrightarrow a} - \mu_{a\leftrightarrow g}\|_2^2 + \|\sigma_{a\leftrightarrow a} - \sigma_{a\leftrightarrow g}\|_2^2 \right) \end{aligned} \tag{14}$$

其中, 均值为 $\mu = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} \boldsymbol{S}^{i,j}$, 标准差 $\sigma = \sqrt{\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\left(\boldsymbol{S}^{i,j} - \mu\right)^2}$, m 和 n 分别为对应相似度矩阵 $\boldsymbol{S}$ 的形状, $\boldsymbol{S}^{i,j}$ 为对应坐标的相似度. 综上所述, 完整的损失函数表达式可记为:

$$\mathcal{L} = (\mathcal{L}_g^c + \mathcal{L}_g^t) + (\mathcal{L}_a^c + \mathcal{L}_a^t) + \lambda_o \mathcal{L}_a^o + \lambda_s \mathcal{L}_a^s \tag{15}$$

其中, λ_o 和 λ_s 为平衡优化目标的超参数.

3　实验

3.1　数据集

所提出的 IDA 框架在当前可用的最大空地行人重识别数据集 CARGO[51] 上进行评估. 该数据集由 13 个摄像机采集得到, 其中 5 个为地面相机视角, 其余 8 个是无人机相机视角. 数据集共收集 5 000 个行人身份, 并按照 1 : 1 的比例划分为训练集和测试集. 训练集中包含 2 500 个行人身份和 51 451 个行人图像; 测试集中包含其余的 2 500 个行人身份和 51 024 个行人图像. 数据集考虑丰富的挑战, 不仅包括显著的空地视角差异难题, 还包括光照、遮挡、分辨率变化等传统重识别难题, 给该任务带来显著的挑战. 在训练集中, 一个行人身份被采集约 10 张行人图像, 平均跨越过 6.47 个地面摄像头和 3.61 个无人机摄像头.

在测试集中, CARGO 共挑选 149 个行人身份作为查询集, 其余测试身份被归类在候选集中作为干扰选项. 测试集中一共设置四种评测协议, 以便充分评测模型在空地场景的性能, 分别为“总体协议” (ALL)、“地−地搜索协议” (G ↔ G)、“空−空搜索协议” (A ↔ A) 和“空−地互搜协议” (A ↔ G). 其中“总体协议”的性能指标是最关键和充分的, 其余

三个协议更关注一些特定模式下的性能. 对于每个查询身份, 在测试集中仅保留两张来自两个不同摄像头下的图像, 一张作为查询, 另一张作为待查询结果留在底库. 在不同的测试评测时, 训练集保持相同.

3.2 实验设置

IDA 的编码器模块采用 ViT-Base[53] 作为骨干网络 (N_e =12), 并在 ImageNet 数据集[57] 上进行预训练. IDA 的解码器模块采用标准解码层堆叠而成, 并随机初始化. IDA 的嵌入特征维度 $d = 768$. 输入图像的大小为 256×128 像素, 并切分为 16×16 像素的图像块, 故 $M = 128$. 在训练阶段构造批数据采样时, 设置 $P = 32$ 并且 $K = 4$. 对于每个输入的图像, 执行概率为 0.5 的随机水平翻转、填充和擦除. IDA 共经过 120 轮训练, 使用随机梯度下降 SGD[58] 作为优化器, 初始学习率为 8×10^{-3}, 经过余弦学习率衰减后, 最终学习率为 1.6×10^{-5}. 使用一种软阈值的方式来实现 $\mathcal{L}_g^t$ 与 $\mathcal{L}_a^t$, 从而避免超参数 α 的选择. 在测试阶段, 不应用任何数据增强策略和重排序技术. 使用 Rank1、mAP 和 mINP 三种指标来评测模型性能检索[2, 18], 其中 1) 首位命中率 Rank1. 旨在衡量查询集中的目标图像对应的真实身份是否在对应检索结果的第一位中被正确识别, 反映模型在最相似结果上的准确度, 该指标越高, 代表模型性能越好. 2) 平均精度均值 mAP. 旨在衡量模型或者方法对于查询集中所有目标图像的平均检索能力, 该指标越高, 代表模型的平均检索性能越好. 3) 反向负惩罚均值 mINP. 用于衡量模型对于最困难相同身份图像的检索能力, 该指标越高, 代表对于困难样本的检索能力越强, 相应模型的性能越好. 具体地, 对于一张查询集中的目标图像, “最困难相同身份图像”被定义为在其对应的检索结果中, 与目标图像具有相同身份且特征相似度最低 (特征距离最远) 的行人图像. 框架基于 PyTorch[20, 59–60] 实现, 并使用一张 GPU 完成实验.

3.3 性能评测

在 CARGO 数据集上的评测结果如表 2 所示, 最佳性能以加粗显示, 将本方法 IDA 框架与几种当前主流的深度行人重识别方法进行对比, 包括 PCB[25–26]、SBS[20]、BoT[33]、MGN[31]、APNet[32]、VV[56]、AGW[2]、TransReID[35] 以及 VDT[51]. 此外, 还采用 ViT-Base 模型[53] 作为基线模型进行比较. 需要指出的是, VDT 被认为是目前最优的空地行人重识别方法, 而 ViT-Base 则可以看作是在未使用摄像头信息训练时, TransReID[35] 的一种简化模型. 为了公平比较, 选择这一模型. 文献 [50] 中的方法并未参与对比, 因为该模型不仅需要行人图像, 还需要额外的行人属性标签作为输入. 由于 CARGO 数据集中并未提供此类属性标签的先验知识, 因此该方法无法直接在 CARGO 数据集上进行评估.

观察表 2, 可以得到如下结论: 1) 所提出的 IDA 模型在多个协议的多个指标上都超过当前最好的 VDT 方法. 在总体指标上, IDA 相较于 VDT 在 Rank1、mAP、mINP 上分别提升 0.32%、2.97%、5.04%, 这表明 IDA 不仅对识别到底库中最相似的行人图像有性能增益, 还对于底库中所有相同身份的行人图像检索都有帮助 (mAP 指标), 尤其是大幅度提高检索到最困难相同身份图像的能力 (mINP 指标), 在“地–地搜索”, “空–空搜索”和“空–地搜索”下分别提高 9.11%, 4.36% 和 3.97%. 这样的性能提

表 2 CARGO 数据集四种协议的主流方法性能评测 (%)

Table 2 The performance evaluation of the mainstream methods for the four protocols of the CARGO dataset (%)

方法	协议 1: ALL			协议 2: G ↔ G			协议 3: A ↔ A			协议 4: A ↔ G		
	Rank1	mAP	mINP	Rank1	mAP	mINP	Rank1	mAP	mINP	Rank1	mAP	mINP
PCB[25–26]	44.23	38.15	26.14	72.32	61.92	45.72	57.50	42.34	22.50	21.25	21.02	14.22
SBS[20]	50.32	43.09	29.76	73.21	62.99	48.24	67.50	49.73	29.32	31.25	29.00	18.71
BoT[33]	54.81	46.49	32.40	77.68	66.47	51.34	65.00	49.79	29.82	36.25	32.56	21.46
MGN[31]	54.49	46.58	33.55	82.14	69.31	53.60	65.00	48.86	27.42	32.50	30.44	21.53
APNet[32]	58.97	50.24	35.76	77.68	66.83	51.85	67.50	54.57	37.35	44.37	39.35	26.76
VV[56]	45.83	38.84	39.57	72.31	62.99	48.24	67.50	49.73	29.32	31.25	29.00	18.71
AGW[2]	60.26	53.44	40.22	81.25	71.66	58.09	67.50	56.48	40.40	43.57	40.90	29.39
TransReID[35]	60.90	53.17	39.57	—	—	—	—	—	—	—	—	—
VDT[51]	64.10	55.20	41.13	82.14	71.59	58.39	82.50	66.83	50.22	48.12	42.76	29.95
基线模型	61.54	53.54	39.62	82.14	71.34	57.55	80.00	64.47	47.07	43.13	40.11	28.20
IDA	**64.42**	**58.17**	**46.17**	**83.04**	**77.04**	**67.50**	**82.50**	**69.65**	**54.58**	**48.75**	**45.13**	**33.92**

升程度也是符合直觉的, 因为在纯地面视角下, 行人的各个部件是相对容易被拍摄到的, 因此局部对齐特征是更容易被匹配到的, 进而性能提升显著. 在无人机视角下, 由于拍摄视角的变化, 行人各部件的拍摄获取难度发生变化, 其中头部等上半身关键部件更容易获取, 腿部等下半身关键部件更难以稳定获取, 因此此时仅有部分局部口令可以被对齐, 跨空地视角下亦同理. 但是本文的方法在面对显著的视角差异引发的行人特征不对齐问题时, 仍然能够通过设计可学习口令特征自发挖掘和关注判别性行人区域, 从而缓解空地行人重识别任务在各种检索模式下的性能表现. 2) 提出的 IDA 模型显著优于已有地面视角的行人重识别方法的对齐策略 (PCB、MGN 和 AGW), 因为这些方法中的采用显式硬行人分块的思想在空地视角下难以生效. 已有分块假设行人是个稳定的刚体, 每个部件出现的位置在地面上相对固定. 然而在无人机视角下, 行人会发生不可控的旋转, 因此行人部件难以通过硬分块策略定位和提取. IDA 则可以通过口令特征提取和隐式对齐局部特征. 同样从表 2 可以看出, 尽管 APNet 和 TransReID 取得相对较好的性能, 但提出的 IDA 在四个协议上均表现出显著优势. 这主要是因为 APNet 与 TransReID 中的模块并非专门针对空地相机场景下的独特性挑战设计, 因此无法有效应对显著的视角差异, 导致行人特征对齐效果不佳, 从而未能实现理想的重识别性能. 相比之下, 本文方法重点针对空地场景的特定挑战, 设计隐式解码对齐的特征学习策略, 从而显著提升重识别性能. 3) IDA 的优越性能并非来自于一个更强大的基线模型, 而是来自于更有效的模型设计. 一方面, 从表 2 中可以看到, 基线模型相比于地面视角方法 (AGW) 和空地视角方法 (VDT) 并不是在所有指标上显著优越的, 然而经过解码器设计和损失函数优化, 其性能提升幅度超过先前的最佳模型性能, 实现优越性能, 说明所提出策略的有效性. 另一方面, 为了更进一步验证方法的优势及其性能增益的来源, 将基线模型与其他经典模型 (如 OSNet[28]、ResNet[29]、ResNeSt[30]) 进行对比, 构造 ViT+OSNet、ViT+ResNet 和 ViT+ResNeSt 作为对比模型, 并在 CARGO 数据集的协议 1 上进行性能对比分析. 如图 3 所示, 当将基线模型与其他模型结合时, 性能并未进一步提升, 反而出现不同程度的性能下降. 这一现象的原因在于, 现有的模型结构并未针对空地相机场景进行专门设计, 因此将这些模型与基线模型结合无法有效解决该场景中面临的巨大视角差异挑战. 更重要的是, 随着引入额外参数的增加 (OSNet → ResNet → ResNeSt), 模型出现

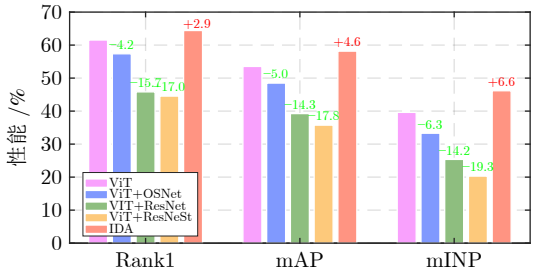


图 3　模型性能增益分析, 其中结果来自 CARGO 数据集的协议 1

Fig. 3　Model performance gain analysis, where results come from Protocol 1 of the CARGO dataset

愈加明显的过拟合问题. 与此相比, 本文所提出的方法是在基线模型的基础上专门针对空地相机场景进行优化设计的, 通过隐式特征对齐机制有效缓解空地视角差异的影响, 从而实现优越的模型性能. 这一结果证明了所提出策略在特定场景下的针对性和有效性.

在 RTX3090 平台上进行基线模型 ViT、VDT 以及本文方法 IDA 的推理计算开销对比实验. 具体来说, 汇报不同模型在 CARGO 数据集中检索一张图像时的推理时间. 实验结果显示, 基线模型 ViT 的推理时间为 1.54 ms/img, VDT 为 1.55 ms/img, 而本文方法 IDA 的推理时间为 1.71 ms/img. 虽然由于 IDA 方法引入额外的解码器结构, 其推理代价略有增加, 但整体上推理时间仍与 ViT 和 VDT 相近, 维持在相同的数量级范围内. 值得注意的是, IDA 方法在模型性能上的提升相对较大, 因此, 所增加的计算开销在可接受的合理范围内, 并在性能提升与计算代价之间实现良好的平衡.

3.4　消融实验

对 IDA 中提出的两个损失函数 $\mathcal{L}_a^o$ 和 $\mathcal{L}_a^s$ 进行消融实验, 以验证其在局部特征学习中的有效性. 其中, 损失函数 $\mathcal{L}_a^o$ 用于约束局部特征提取, 使其学习到彼此不重叠的行人判别区域, 而 $\mathcal{L}_a^s$ 则旨在约束视角内和视角间的局部特征分布趋于一致.

实验结果如表 3 所示, 总结如下: 1) 在总体协议中, 在基础模型 $\mathcal{F}$ 的基础上引入这两个损失函数后, 性能得到显著提升. 具体而言, 在不添加任何损失函数的情况下, 所提出模型 $\mathcal{F}$ 的 Rank1、mAP 和 mINP 分别为 60.26%、53.89% 和 41.36%. 仅单独加入 $\mathcal{L}_a^o$ 后, 这些指标分别提高 3.52%、3.66% 和 4.33%; 单独加入 $\mathcal{L}_a^s$ 则分别带来 0.96%、1.87% 和 2.86% 的增益. 当两者同时引入时, 性能进一步提升, 分别达到 4.16%、4.28% 和 4.81%, 这表明了这两个损失函数的互补性及其在特征学习中的增益效

表 3 IDA 框架消融实验 (%)

Table 3 Ablation study of IDA framework (%)

			协议 1: ALL			协议 2: G ↔ G			协议 3: A ↔ A			协议 4: A ↔ G		
$\mathcal{F}(\cdot)$	$\mathcal{L}_a^o$	$\mathcal{L}_a^s$	Rank1	mAP	mINP	Rank1	mAP	mINP	Rank1	mAP	mINP	Rank1	mAP	mINP
✓			60.26	53.89	41.36	81.25	73.66	62.70	75.00	63.94	47.10	43.75	40.33	28.71
✓	✓		63.78	57.55	45.69	**83.93**	**77.33**	**68.21**	77.50	64.55	47.52	46.25	43.97	32.88
✓		✓	61.22	55.76	44.22	82.14	75.53	66.28	80.00	69.36	55.21	44.37	41.82	30.88
✓	✓	✓	**64.42**	**58.17**	**46.17**	83.04	77.04	67.50	**82.50**	**69.65**	**54.58**	**48.75**	**45.13**	**33.92**

果. 2) 在四种协议下, $\mathcal{L}_a^o + \mathcal{L}_a^s$ 的组合在大多数指标上都实现性能提升. 尽管在“地−地”搜索协议中, $\mathcal{L}_a^o + \mathcal{L}_a^s$ 略低于单独使用 $\mathcal{L}_a^o$ 的效果, 但依然达到相当接近的可比较性能水平. 综合所有协议的整体表现, 特别是在 ALL 协议下的结果, 可以认为 $\mathcal{L}_a^o$ 和 $\mathcal{L}_a^s$ 的组合是实现最佳模型性能的有效策略.

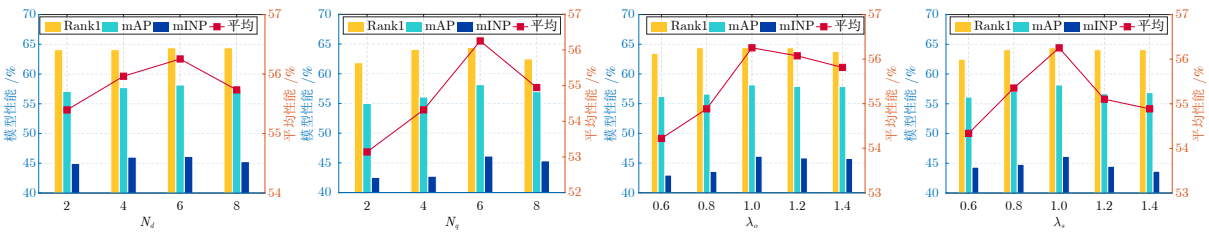
通过分析, $\mathcal{L}_a^o$ 在“地−地”协议下实现良好性能的原因在于, 该协议中的行人图像仅包含地面视角图像, 具有良好的潜在对齐性, 在这种情况下, 仅学习判别性局部特征是足够的. 然而在其他协议下, 这种对齐性被破坏, 因此还需要 $\mathcal{L}_a^s$ 的参与以进一步调整和优化特征学习.

3.5 参数分析

在 IDA 框架中, 引入四个需要手动调整的参数, 分别为 N_d、N_q、λ_o 和 λ_s. 其中, N_d 和 N_q 与模型结构设计相关, 而 λ_o 和 λ_s 则影响优化目标的权重. 在本节中, 基于模型在总体协议下的性能表现, 对这些参数进行分析. 具体地, 当对某一个参数进行分析时, 其余参数均维持最佳参数保持不变, 从而更加直观地分析某一参数对于模型性能水平的影响, 其中平均性能代表三个指标的平均值, 便于直观展示最佳参数的选择. 并将模型在 Rank1、mAP 和 mINP 三个指标上的平均性能作为参数选择的依据, 这是由于, 一方面采用平均指标性能挑选最佳模型, 也是行人重识别领域中一种常见的模型挑选策略 (例如 FastReID[20]); 除此之外, 尽管三个指标是从不同的角度描述模型的性能, 但是所用的三个指标均是单调递增指标 (即指标数值越高, 模型的性能越好), 因此, 三个指标的平均值越高, 也可以在一定程度上反映模型的三个指标同时实现较高的数值. 具体实验结果展示在图 4 中, 以下是逐一分析.

参数 N_d 分析. N_d 控制着解码模块 $\mathcal{F}_d$ 中的解码层数. 分别挑选 {2, 4, 6, 8} 作为候选参数, 结果如图 4(a) 所示. 可以看到, 随着 N_d 的增大, 模型性能呈现出先增大后减少的趋势, 并在 $N_d = 6$ 时取得最佳性能. 可以看出, 当 N_d 比较小时, 此时解码模块 $\mathcal{F}_d$ 比较浅, 较少的参数空间使得模型无法充分学习到局部特征表达; 当 N_d 比较大时, 此时解码模块 $\mathcal{F}_d$ 比较深, 大量的参数学习使得模型极易过拟合, 从而实现次优的模型性能.

参数 N_q 分析. N_q 控制着解码模块 $\mathcal{F}_d$ 中的局部对齐口令特征 $[t_a^1, \cdots, t_a^{N_q}]$ 的数量, 在损失函数 $\mathcal{L}_a^o$ 的作用下, N_q 也进一步代表着模型可学习的行人局部部件区域个数. 分别挑选 {2, 4, 6, 8} 作为候选参数, 结果如图 4(b) 所示. 可以看到, 模型性能先随着 N_q 的增大而增大, 在 $N_q = 6$ 时取得最佳性能, 并在随后而减少. 可以看出, 当 N_q 比较小时, 此时局部口令的数量相对较少, 对于行人的判别性部件区域无法充分的挖掘和捕捉, 从而实现次优的性能; 当 N_q 比较大时, 此时解码模块中的可学习局部口令数量过多, 在正交性的约束下使得部分口令过拟合关注与身份无关的背景信息, 导致模型性能



(a) 参数 N_d (b) 参数 N_q (c) 参数 λ_o (d) 参数 λ_s

(a) Parameter N_d (b) Parameter N_q (c) Parameter λ_o (d) Parameter λ_s

图 4 IDA 框架中的参数分析

Fig. 4 Parameter analysis in the IDA framework

的下降.

参数 λ_o 分析. λ_o 控制整体优化目标 $\mathcal{L}$ 中正交性损失函数 $\mathcal{L}_a^o$ 所占的比重. 分别挑选 {0.6, 0.8, 1.0, 1.2, 1.4} 作为候选参数, 结果如图 4(c) 所示. 可以看到, 模型性能先随着 λ_o 的增大而增大, 在 $\lambda_o = 1.0$ 时取得最佳性能, 并在之后随着 λ_o 的进一步增大而减少. 可以看出, 当 λ_o 比较小时, 对于局部口令关注不相关的局部区域是松弛的, 导致局部口令特征之间存在较大的冗余, 从而导致次优的表征学习; 当 λ_o 比较大时, 优化目标过分关注局部特征之间的不相关性, 从而轻视对于局部特征身份判别性的约束, 导致模型的性能下降.

参数 λ_s 分析. λ_s 控制整体优化目标 $\mathcal{L}$ 中一致性损失函数 $\mathcal{L}_a^s$ 所占的比重. 分别挑选 {0.6, 0.8, 1.0, 1.2, 1.4} 作为候选参数, 结果如图 4(d) 所示. 可以看到, 模型性能先随着 λ_s 的增大而增大, 在 $\lambda_s = 1.0$ 时取得最佳性能, 并在之后会降低. 可以看出, 当 λ_s 比较小时, 模型整体不关注视角内以及视角间的一致性特征分布, 致使特征之间的视角差异性偏置没有被良好的缓解, 从而导致次优的模型性能; 当 λ_s 比较大时, 优化目标以特征的一致性分布为主要优化方向, 从而轻视对于其他优化目标的学习, 例如身份相关的优化目标, 一个欠佳的优化目标平衡导致模型的性能下降.

3.6　可视化分析

在图 5 中详细展示本文提出的 IDA 模型在 CARGO 数据集四个评测协议下的前五名 (Rank 1-5) 可视化检索结果, 并与基线模型的结果进行对比, 其中红色框代表正确检索结果.

可以看到在四种协议下, 本文方法均实现比基线模型更优越的检索性能, 具体体现为成功检索到更多的正确行人图像 (在“地−地”搜索协议中), 或者实现精度更高的检索序列 (在总体协议、“空−空”搜索以及“空−地”互搜协议中). 通过观察检索结果可发现, 基线模型由于缺乏对于局部特征的关注, 在多种协议下对于表观相近的困难行人无法实现良好的区分, 从而导致检索的失败. 例如在整体协议中, 查询图像的行人裤子与基线模型首位命中的行人存在显著不同, 但是这被基线模型所忽略, 本文方法对于局部特征有着显式的关注, 能够实现在跨视角挑战下的学习和对齐, 因此可以对该困难样本实现良好的区分.

在其他三种协议下也是类似的情况, 例如“地−地”搜索协议中待查询行人的头发 (图 5(b))、“空−地”互搜协议中待查询行人的鞋子颜色 (图 5(d)), 本文所提出的方法都可以关注到这些具有判别性的局部特征, 从而返回一个正确的检索结果.

4　结束语

在本文中, 关注跨空地行人重识别这一任务, 并提出基于隐式解码特征对齐的特征提取框架以缓解该场景下显著的视角差异挑战. 具体来说, 首先设计基于解码对齐的自注意力模型, 通过在自注意

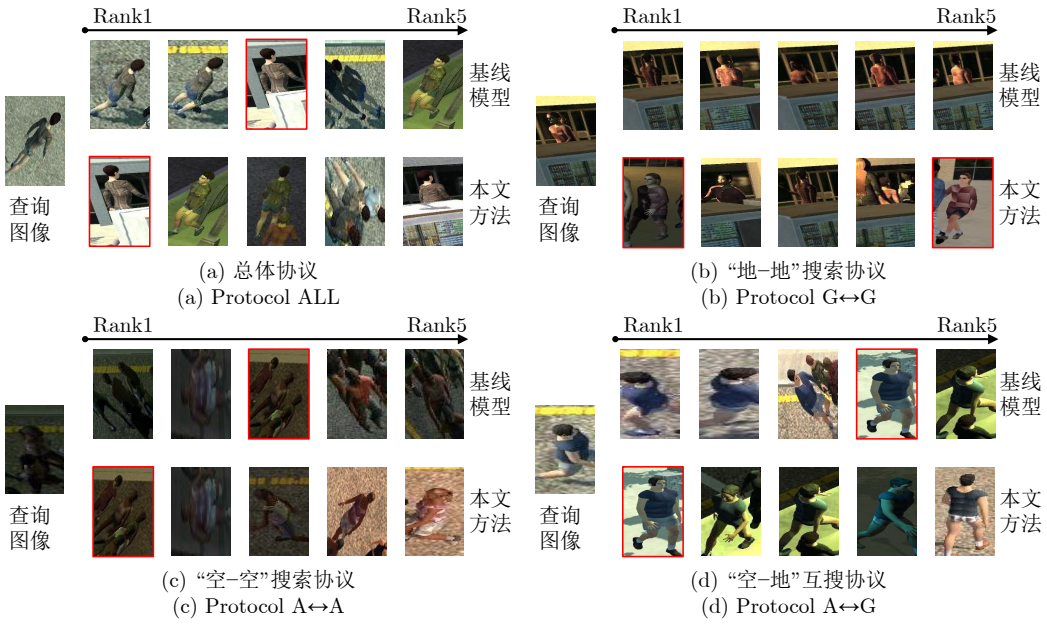


图 5　IDA 框架在 CARGO 数据集四种协议下的检索可视化分析

Fig. 5　Retrieval visualization of IDA framework under the four protocols of the CARGO dataset

力编码器的基础上引入解码模块和一系列的可学习口令特征自主挖掘和提取判别性行人局部特征, 实现具有判别性的行人表征学习. 其次, 设计正交损失函数和一致性损失函数来进一步指导特征的学习与分布. 最后, 在 CARGO 数据集上评估方法, 结果表明 IDA 框架实现了当前最优的模型性能, 所提出的隐式对齐策略显著超越已有的重识别方法.

此外, 也必须指出, 尽管本文提出的方法能够显著提升在具有大视角差异场景中的表现, 但在空地视角差异不显著的场景 (例如 AG-ReID 数据集) 中, 性能提升存在一定限制. 在这些场景中, 方法的潜力未能完全发挥. 针对这一局限性, 计划在未来的工作中进一步研究并扩展方法的应用场景与泛化能力, 以更好地应对更广泛的任务需求.

References

1 Ye Yu, Wang Zheng, Liang Chao, Han Zhen, Chen Jun, Hu Rui-Min. A survey on multi-source person re-identification. *Acta Automatica Sinica*, 2020, **46**(9): 1869−1884
(叶钰, 王正, 梁超, 韩镇, 陈军, 胡瑞敏. 多源数据行人重识别研究综述. 自动化学报, 2020, **46**(9): 1869−1884)

2 Ye M, Chen S Y, Li C Y, Zheng W S, Crandall D, Du B. Transformer for object re-identification: A survey. *International Journal of Computer Vision*, 2025, **133**(5): 2410−2440

3 Zhang Q, Lai J H, Xie X H, Jin X F, Huang S E. Separable spatial-temporal residual graph for cloth-changing group re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, **46**(8): 5791−5805

4 Zhang Q, Lai J H, Feng Z X, Xie X H. Uncertainty modeling for group re-identification. *International Journal of Computer Vision*, 2024, **132**(8): 3046−3066

5 Luo Hao, Jiang Wei, Fan Xing, Zhang Si-Peng. A survey on deep learning based person re-identification. *Acta Automatica Sinica*, 2019, **45**(11): 2032−2049
(罗浩, 姜伟, 范星, 张思朋. 基于深度学习的行人重识别研究进展. 自动化学报, 2019, **45**(11): 2032−2049)

6 Zhang Yong-Fei, Yang Hang-Yuan, Zhang Yu-Jia, Dou Zhao-Peng, Liao Sheng-Cai, Zheng Wei-Shi, et al. Recent progress in person re-ID. *Journal of Image and Graphics*, 2023, **28**(6): 1829−1862
(张永飞, 杨航远, 张雨佳, 豆朝鹏, 廖胜才, 郑伟诗, 等. 行人再识别技术研究进展. 中国图象图形学报, 2023, **28**(6): 1829−1862)

7 Sarker P K, Zhao Q J, Uddin M K. Transformer-based person re-identification: A comprehensive review. *IEEE Transactions on Intelligent Vehicles*, 2024, **9**(7): 5222−5239

8 Yang X, Liu H L, Wang N N, Gao X B. Image-level adaptive adversarial ranking for person re-identification. *IEEE Transactions on Image Processing*, 2024, **33**: 5172−5182

9 Cui Z Y, Zhou J H, Peng Y X. DMA: Dual modality-aware alignment for visible-infrared person re-identification. *IEEE Transactions on Information Forensics and Security*, 2024, **19**: 2696−2708

10 He W Z, Deng Y H, Tang S X, Chen Q H, Xie Q S, Wang Y Z, et al. Instruct-ReID: A multi-purpose person re-identification task with instructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2024. 17521−17531

11 Ye M, Shen W, Zhang J W, Yang Y, Du B. SecureReID: Privacy-preserving anonymization for person re-identification. *IEEE Transactions on Information Forensics and Security*, 2024, **19**: 2840−2853

12 Zhang Q, Lai J H, Feng Z X, Xie X H. Seeing like a human: Asynchronous learning with dynamic progressive refinement for person re-identification. *IEEE Transactions on Image Processing*, 2022, **31**: 352−365

13 Nguyen H, Nguyen K, Sridharan S, Fookes C. AG-ReID.v2: Bridging aerial and ground views for person re-identification. *IEEE Transactions on Information Forensics and Security*, 2024, **19**: 2896−2908

14 Wang L, Zhang Q, Qiu J Y, Lai J H. Rotation exploration transformer for aerial person re-identification. In: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME). Niagara Falls, Canada: IEEE, 2024. 1−6

15 Qiu J Y, Feng Z X, Wang L, Lai J H. Salient part-aligned and keypoint disentangling transformer for person re-identification in aerial imagery. In: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME). Niagara Falls, Canada: IEEE, 2024. 1−6

16 Khaldi K, Nguyen V D, Mantini P, Shah S. Unsupervised person re-identification in aerial imagery. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW). Waikoloa, USA: IEEE, 2024. 260−269

17 Ding J, Xue N, Xia G S, Bai X, Yang W, Yang M Y, et al. Object detection in aerial images: A large-scale benchmark and challenges. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022, **44**(11): 7778−7796

18 Ye M, Shen J B, Lin G J, Xiang T, Shao L, Hoi S C H. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022, **44**(6): 2872−2893

19 Zhu K, Guo H, Zhang S, Wang Y, Liu J, Wang J, et al. AAformer: Auto-aligned transformer for person re-identification. *IEEE Transactions on Neural Networks and Learning Systems*, 2024, **35**(12): 17307−17317

20 He L X, Liao X Y, Liu W, Liu X C, Cheng P, Mei T. FastReID: A PyTorch toolbox for general instance re-identification. In: Proceedings of the 31st ACM International Conference on Multimedia. Ottawa, Canada: Association for Computing Machinery, 2023. 9664−9667

21 Farenzena M, Bazzani L, Perina A, Murino V, Cristani M. Person re-identification by symmetry-driven accumulation of local features. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco, USA: IEEE, 2010. 2360−2367

22 Gheissari N, Sebastian T B, Hartley R. Person reidentification using spatiotemporal appearance. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). New York, USA: IEEE, 2006. 1528−1535

23 Layne R, Hospedales T M, Gong S G. Person re-identification by attributes. In: Proceedings of the British Machine Vision Conference. Surrey, UK: BMVC, 2012. Article No. 24

24 Liu C X, Gong S G, Loy C C, Lin X G. Person re-identification: What features are important? *Computer Vision Workshops-ECCV 2012. Workshops and Demonstrations*. Berlin, Heidelberg: Springer, 2012. 391−401

25 Sun Y F, Zheng L, Yang Y, Tian Q, Wang S J. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In: Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer, 2018. 501−518

26 Sun Y F, Zheng L, Li Y L, Yang Y, Tian Q, Wang S J. Learning part-based convolutional features for person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021, **43**(3): 902−917

27 Zhou K Y, Yang Y X, Cavallaro A, Xiang T. Omni-scale feature learning for person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea: IEEE, 2019. 3701−3711

28 Zhou K Y, Yang Y X, Cavallaro A, Xiang T. Learning generalisable omni-scale representations for person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022, **44**(9): 5056−5069

29 He K M, Zhang X Y, Ren S Q, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016. 770−778

30 Zhang H, Wu C R, Zhang Z Y, Zhu Y, Lin H B, Zhang Z, et al. ResNeSt: Split-attention networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). New Orleans, USA: IEEE, 2022. 2735−2745

31 Wang G S, Yuan Y F, Chen X, Li J W, Zhou X. Learning discriminative features with multiple granularities for person re-identification. In: Proceedings of the 26th ACM International Conference on Multimedia. Seoul, The Republic of Korea: Association for Computing Machinery, 2018. 274−282

32 Chen G Y, Gu T P, Lu J W, Bao J A, Zhou J. Person re-identification via attention pyramid. *IEEE Transactions on Image Processing*, 2021, **30**: 7663−7676

33 Luo H, Gu Y Z, Liao X Y, Lai S Q, Jiang W. Bag of tricks and a strong baseline for deep person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Long Beach, USA: IEEE, 2019. 1487−1495

34 Luo H, Jiang W, Zhang X, Fan X, Qian J J, Zhang C. AlignedReID++: Dynamically matching local information for person re-identification. *Pattern Recognition*, 2019, **94**: 53−61

35 He S T, Luo H, Wang P C, Wang F, Li H, Jiang W. TransReID: Transformer-based object re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE, 2021. 14993−15002

36 Yang Q, Chen Y K, Peng D Z, Peng X, Zhou J T, Hu P. Noisy-correspondence learning for text-to-image person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2024. 27187−27196

37 Zhu J Q, Wu H X, Chen Y T, Xu H, Fu Y Q, Zeng H Q, et al. Cross-modal group-relation optimization for visible-infrared person re-identification. *Neural Networks*, 2024, **179**: Article No. 106576

38 Zhu H D, Budhwant P, Zheng Z H, Nevatia R. SEAS: ShapE-Aligned supervision for person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2024. 164−174

39 Bai S T, Chang H, Ma B P. Incorporating texture and silhouette for video-based person re-identification. *Pattern Recognition*, 2024, **156**: Article No. 110759

40 Yu Z Y, Li L S, Xie J L, Wang C S, Li W J, Ning X. Pedestrian 3D shape understanding for person re-identification via multi-view learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024, **34**(7): 5589−5602

41 Yang B, Chen J, Ye M. Shallow-deep collaborative learning for unsupervised visible-infrared person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2024. 16870−16879

42 Nguyen K, Fookes C, Sridharan S, Liu F, Liu X M, Ross A, et al. AG-ReID 2023: Aerial-ground person re-identification challenge results. In: Proceedings of 2023 IEEE International Joint Conference on Biometrics (IJCB). Ljubljana, Slovenia: IEEE, 2023. 1−10

43 Liu M, Wang F, Wang X P, Wang Y N, Roy-Chowdhury A K. A two-stage noise-tolerant paradigm for label corrupted person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, **46**(7): 4944−4956

44 Huang Y, Wu Q, Zhang Z, Shan C F, Huang Y, Zhong Y, et al. Meta clothing status calibration for long-term person re-identification. *IEEE Transactions on Image Processing*, 2024, **33**: 2334−2346

45 Rami H, Giraldo J H, Winckler N, Lathuilière S. Source-guided similarity preservation for online person re-identification. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Waikoloa, USA: IEEE, 2024. 1700−1709

46 Nguyen V D, Mantini P, Shah S K. Temporal 3D shape modeling for video-based cloth-changing person re-identification. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW). Waikoloa, USA: IEEE, 2024. 173−182

47 Liu X H, Zhang P P, Yu C Y, Qian X S, Yang X Y, Lu H C. A video is worth three views: Trigeminal transformers for video-based person re-identification. *IEEE Transactions on Intelligent Transportation Systems*, 2024, **25**(9): 12818−12828

48 Zhang S Z, Zhang Q, Yang Y F, Wei X, Wang P, Jiao B L, et al. Person re-identification in aerial imagery. *IEEE Transactions on Multimedia*, 2021, **23**: 281−291

49 Li T J, Liu J, Zhang W, Ni Y, Wang W Q, Li Z H. UAV-human: A large benchmark for human behavior understanding with unmanned aerial vehicles. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2021. 16261−16270

50 Nguyen H, Nguyen K, Sridharan S, Fookes C. Aerial-ground person re-ID. In: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME). Brisbane, Australia: IEEE, 2023. 2585−2590

51 Zhang Q, Wang L, Patel V M, Xie X H, Lai J H. View-decoupled transformer for person re-identification under aerial-ground camera network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2024. 22000−22009

52 He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016. 770−778

53 Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, et al. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc., 2017. 6000−6010

54 Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. In: Proceedings of the 9th International Conference on Learning Representations. ICLR, 2021. 1−21

55 Wu J B, Liu H, Su Y X, Shi W, Tang H. Learning concordant attention via target-aware alignment for visible-infrared person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE, 2023.

11088−11097

56 Chen S Y, Ye M, Du B. Rotation invariant transformer for recognizing object in UAVs. In: Proceedings of the 30th ACM International Conference on Multimedia. New York, USA: Association for Computing Machinery, 2022. 2565−2574

57 Kuma R, Weill E, Aghdasi F, Sriram P. Vehicle re-identification: An efficient baseline using triplet embedding. In: Proceedings of the International Joint Conference on Neural Networks (IJCNN). Budapest, Hungary: IEEE, 2019. 1−9

58 Deng J, Dong W, Socher R, Li L J, Li K, Li F F. ImageNet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA: IEEE, 2009. 248−255

59 Robbins H, Monro S. A stochastic approximation method. *The Annals of Mathematical Statistics*, 1951, **22**(3): 400−407

60 Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An imperative style, high-performance deep learning library. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2019. Article No. 721

贝俊仁 中山大学计算机学院硕士研究生. 主要研究方向为行人重识别与计算机视觉.
E-mail: beijr@mail2.sysu.edu.cn
(**BEI Jun-Ren** Master student at the School of Computer Science and Engineering, Sun Yat-sen University. His research interest covers person re-identification and computer vision.)

张 权 中山大学系统科学与工程学院博士后. 主要研究方向为行人重识别与计算机视觉. 本文通信作者.
E-mail: zhangq689@mail.sysu.edu.cn
(**ZHANG Quan** Postdoctor at the School of Systems Science and Engineering, Sun Yat-sen University. His research interest covers person re-identification and computer vision. Corresponding author of this paper.)

赖剑煌 中山大学计算机学院教授. 主要研究方向为计算机视觉与模式识别. E-mail: stsljh@mail.sysu.edu.cn
(**LAI Jian-Huang** Professor at the School of Computer Science and Engineering, Sun Yat-sen University. His research interest covers computer vision and pattern recognition.)