http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2025.0834

基于大模型属性解析的空地行人检索方法

谭荃戈[1]，王蓉[1]，廖曼丞[2]，李欣[1,*]
(1. 中国人民公安大学 信息网络安全学院，北京 100038；2. 中国人民公安大学 侦查学院，北京 100038)

摘 要：可在监控探头分布稀疏区域协同分析无人机拍摄图像与固定监控视频图像，基于行人图像的跨视角身份认定实现对目标持续追踪的巡检任务。无人机俯视视角与地面监控的平视视角存在显著差异，导致基于传统监控场景设计的图像检索算法在地空跨视角场景中迁移性能受限。现有空地行人检索方法主要聚焦于缓解跨视角引起的表观差异，而对行人属性特征的挖掘与分析尚未充分深入。为解决上述问题，提出基于大模型属性解析的空地行人检索方法。基于多模态大模型构建行人解析模块，生成细粒度语义属性并设计属性三元组损失实现跨视角语义对齐；引入视角解耦架构，通过分层减法分离剥离视角相关特征，并施加正交损失约束特征独立性；引入多尺度扩增Transformer，结合多尺度扩张注意力与全局自注意力优化计算复杂度与感受野平衡，降低模型参数量。在AG-ReID.v1、AG-ReID.v2、CARGO数据集上的实验表明：所提方法能够有效提升空地行人检索任务在Rank-1、mAP与mINP指标上的性能，证明了该方法的有效性。

关 键 词：UAV；行人检索；空地图像；属性解析；多模态大语言模型
中图分类号：TP391.4; V279
文献标志码：A **文章编号**：1001-5965(2026)09-3108-09

无人机巡检技术的发展为现代安防技术体系提供了新方法。无人机俯视角相机可快速覆盖传统地面监控无法企及的区域，实时捕捉人员动态。地空行人检索技术能够将航拍行人目标的视觉特征与地面监控行人特征进行跨视角关联匹配，在监控探头稀疏区域实现无人机与固定监控探头协同的行人持续搜索与轨迹追踪。

经典的行人检索方法多基于卷积神经网络构建，基于部件的卷积基线(part-based convolutional baseline, PCB)网络[1]、多粒度网络(multiple granularity network, MGN)[2]通过多分支结构和细粒度特征学习显著提升性能，文献[1,3]通过融入注意力机制增强行人特征表征能力，在Market1501[4]、MSMT17多场景多时段行人重识别数据集[5]等地面固定视角数据集上取得突破。然而，卷积神经网络的有限感受野难以应对复杂场景的时空关联建模，这推动Transformer[6]架构在行人检索领域的应用：基于Transformer的目标重识别模型(Transformer-based object re-identification, TransReID)[7]及其衍生模型[8-11]通过全局自注意力机制捕获特征长距离依赖，在遮挡和姿态变化场景表现出色。尽管这些行人检索方法能够对不同摄像头视角下的同一行人进行有效检索，但其固定部署模式仍无法应对稀疏摄像头覆盖区域的目标搜索需求，因此，无人机的普及催生空地协同行人检索[12]研究。

空地行人检索作为新兴方向，致力于解决无人机俯视视角与地面监控视角间的跨平台身份关联难题。该领域面临双重挑战：首先，视角差异引发的人体几何畸变，使经典行人检索模型的跨视角匹配精度下降；其次，无人机影像分辨率较低，导致细粒

收稿日期：2025-12-02；录用日期：2026-01-25；网络出版时间：2026-04-03 10:45
网络出版地址：link.cnki.net/urlid/11.2625.V.20260402.1729.004
基金项目：国家自然科学基金(62076246)
*通信作者. E-mail：lixin@ppsuc.edu.cn

引用格式：谭荃戈，王蓉，廖曼丞，等. 基于大模型属性解析的空地行人检索方法[J]. 北京航空航天大学学报，2026，52(9)：3108-3116.
Tan Q G，Wang R，Liao M C，et al. Large model attribute parsing-based aerial-ground pedestrian retrieval method[J]. Journal of Beijing University of Aeronautics and Astronautics，2026，52(9)：3108-3116 (in Chinese).

度特征提取困难。Schumann 等[12] 构建首个空地行人检索数据集并提出一种基于深度学习的行人检索方法。Nguyen 等[13] 提出一种基于软属性解释的双流模型, 并在后续工作中扩展为可解释三流模型[14]。Zhang 等[15] 设计视角解耦 Transformer 网络从全局特征中分离视角特征, 并通过 3D 模拟创建了 CARGO 数据集。Wang 等[16] 通过引入扩散模型生成不同视角之间的视图特定特征来检索目标人员。上述方法主要针对地空视角的行人视觉特征差异进行分析, 其中, 基于属性的方法是通过模型获取的“软属性”, 需要人工对图像中的每个行人进行属性标注, 并且训练模型预测行人属性。属性行人检索通过引入语义先验突破细粒度识别瓶颈。渐进式注意力增强网络(progressive attention-enhanced network, PAENet)[17] 引入了 8 个属性作为注释来学习细粒度的语义属性信息。Tan 等[18] 设计了多个模板指导多模态大模型生成图像标签用于模型训练。Yang 等[19] 利用扩散模型构建了包含用于行人检索的多属性与语言搜索(multi-attribute and language search, MALS)数据集, 并同时进行了文本匹配学习和属性学习。这些方法已经证明引入属性信息对于地面监控视角下的行人目标身份认定具有优势。

上述方法已经证明已有工作有效提升了模型在空对地场景中的特征判别能力、泛化能力、对行人外观显著变化的适应能力。然而, 低空无人机视角与地面监控视角之间存在显著差异, 使得模型容易忽略对行人检索至关重要的细节特征, 制约了检索精度的提升。为挖掘行人细节特征, 已有方法通过分析行人属性提升模型细节特征判别力, 但应用于地空行人检索方法中的属性需要人工进行标注, 且不同视角图像的造成属性预测准确度较低。因此, 本文通过引入多模态大模型进行空地行人属性解析与属性三元组损失约束模型学习一致性的身份特征, 通过属性描述语义缓解视角差异带来的影响, 减少对大规模人工属性标注的依赖。

本文的主要创新点如下:

1) 提出多模态零样本属性解析机制, 利用多模态大模型解析空地行人图像, 构建属性三元组损失。

2) 引入视角解耦架构, 有效分离视角相关/无关特征。

3) 引入多尺度扩增 Transformer[20], 通过多尺度扩增注意力(multi-scale dilated attention, MSDA)与多头自注意力(multi-head self-attention, MHSA)模块级联优化计算, 降低计算参数量, 提升模型运行速度。

1　行人属性解析空地行人检索方法

1.1　本文方法总体架构

本文提出一种基于大模型属性解析的空地行人检索方法, 该方法的整体框架如图 1 所示。模型采用双分支结构, 分别处理地面监控视角和无人机航拍视角的输入图像。首先, 基于多模态大模型的属性解析模块对两类视角图像进行细粒度语义属性(如衣着颜色、背包类型等)的联合提取与对齐, 并构建属性三元组损失以增强跨视角属性一致性。在此基础上, 网络通过分层减法分离单元正交损失函数组成的视角解耦架构, 显式分离视角相关特征与视角无关特征, 从而抑制视角变异对身份判别的影响。特征提取主干采用多尺度扩增 Transformer 网络, 其低层网络通过 MSDA 块捕获局部细节特征, 而高层网络利用全局自注意力机制建模长程依赖关系, 在计算效率与特征表达能力之间取得平衡。最终, 通过融合属性监督、视角解耦约束及多尺度特征表示, 模型能够在跨视角、低分辨率的无人机巡检场景下实现鲁棒的行人检索。

1.2　多模态大模型属性解析

针对地空跨视角行人检索中存在的视角差异大、类内变化显著等问题, 本文提出基于多模态大模型的属性解析方法, 通过编码结构化的属性特征引入语义信息, 引导相似语义的行人图像缩小特征间距离。利用 GLM-4V-Flash[21] 的强大语义理解能力提取结构化属性特征, 并结合属性三元组损失增强模型对判别性属性的学习。

首先, 给定输入图像 $I \in \mathbf{R}^{H\times W\times 3}$, 采用 GLM-4V-Flash 模型进行细粒度属性解析。该模型通过视觉-语言预训练具备优异的跨模态理解能力, 可输出 9 类结构化属性描述, 包括性别及上衣、下衣、鞋、配饰的类别与颜色:

$$A = \{\boldsymbol{a}_{\text{gender}}, \boldsymbol{a}_{\text{upper_type}}, \cdots, \boldsymbol{a}_{\text{accessory_color}}\} \tag{1}$$

为将离散属性转化为连续的嵌入特征, 本文设计了分层属性编码架构。首先, 为每个属性类别建立独立的嵌入矩阵 $\boldsymbol{E}_i \in \mathbf{R}^{d_i\times k_i}$ (k_i 为第 i 类属性的取值数量), 将各属性 $\boldsymbol{a}_i$ 映射为嵌入向量 $\boldsymbol{e}_i = \boldsymbol{E}_i \cdot \boldsymbol{a}_i$。随后将全部 9 类属性嵌入拼接为复合向量:

$$\boldsymbol{e}_{\text{concat}} = \text{Concat}(\boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_9) \tag{2}$$

并通过全连接层 $\boldsymbol{W}_f \in \mathbf{R}^{d_a\times \sum d_i}$ 与 ReLU 激活函数生成最终属性特征:

$$\boldsymbol{f}_{\text{attr}} = \text{ReLU}(\boldsymbol{W}_f \cdot \boldsymbol{e}_{\text{concat}}) \tag{3}$$

为进一步增强属性特征的判别能力, 设计属性

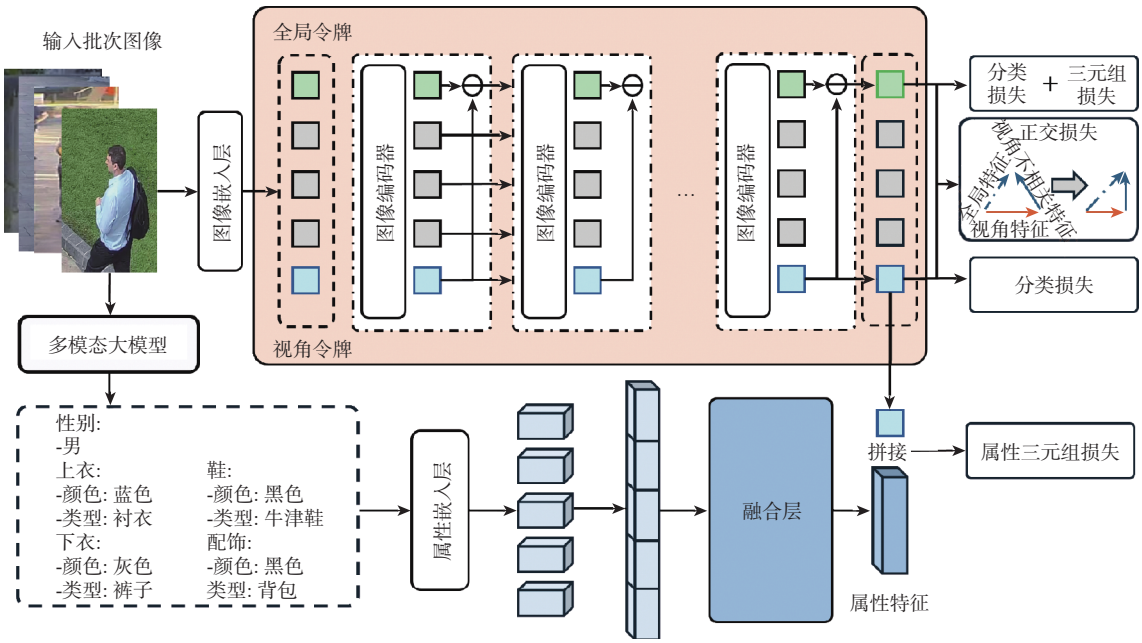


图 1 本文方法整体框架

Fig. 1 Overall framework of the proposed method

指导的三元组损失。该方法将全局特征 $\boldsymbol{f}_{\text{global}}$ 与属性特征 $\boldsymbol{f}_{\text{attr}}$ 拼接为融合特征：

$$\tilde{\boldsymbol{f}} = [\boldsymbol{f}_{\text{global}} \| \boldsymbol{f}_{\text{attr}}] \tag{4}$$

在此基础上计算改进的三元组损失：

$$L_{\text{attr}} = \max(0, D(\tilde{\boldsymbol{f}}_{\text{a}}, \tilde{\boldsymbol{f}}_{\text{p}}) - D(\tilde{\boldsymbol{f}}_{\text{a}}, \tilde{\boldsymbol{f}}_{\text{n}}) + \alpha) \tag{5}$$

式中：$\tilde{\boldsymbol{f}}_{\text{a}}$ 表示锚点样本特征；$\tilde{\boldsymbol{f}}_{\text{p}}$ 表示正样本特征；$\tilde{\boldsymbol{f}}_{\text{n}}$ 表示负样本特征；$D(\cdot)$ 采用余弦距离度量；边界值 α 根据正负样本的属性相似度动态调整：

$$\alpha = \alpha_{\text{base}} + \lambda_{\alpha}(1 - \text{sim}(\boldsymbol{A}_{\text{p}}, \boldsymbol{A}_{\text{n}})) \tag{6}$$

其中：α_{base} 表示基准边界值；$\boldsymbol{A}_{\text{p}}$ 表示正样本属性；$\boldsymbol{A}_{\text{n}}$ 表示负样本属性；λ_{α} 表示平衡权重；sim(·)计算两样本属性的 Jaccard 相似度；λ 用于控制属性相似度对边界的影响程度。该机制使模型对属性差异小的困难样本对施加更严格的约束，从而提升特征判别性。

1.3 视角解耦架构

为应对无人机巡检中地面平视视角与高空俯视视角间的剧烈差异对行人特征表达的干扰，本文引入视角解耦架构，其核心是通过结构化的特征分离机制显式解耦视角相关特征与视角无关的身份特征。该架构以 Transformer 为基础框架，融入了 2 个可学习令牌——全局令牌 $\boldsymbol{t}_{\text{m}}$ 用于聚合全局身份语义，视角令牌 $\boldsymbol{t}_{\text{v}}$ 专用于捕捉拍摄视角特性（如俯视造成的形变、地面视角的姿态变化）。在特征处理过程中，核心在于分层减法分离机制的设计：输入令牌序列经多层 Transformer 块逐级处理，每层首先通过自注意力操作更新包括 $\boldsymbol{t}_{\text{m}}$ 和 $\boldsymbol{t}_{\text{v}}$ 在内的所有令牌表示；接着，执行特征解耦操作，即从更新后的元令牌表示中显式减去对应的视角令牌表示：

$$\boldsymbol{t}_{\text{m}}^{(j+1)} = \boldsymbol{t}_{\text{m}}^{(j+1)} - \boldsymbol{t}_{\text{v}}^{(j+1)} \tag{7}$$

这种层级递进的减法策略使网络在特征学习过程中持续剥离身份表示中的视角噪声，使深层 $\boldsymbol{t}_{\text{m}}$ 愈发聚焦于与视角无关的鲁棒行人特征。

为确保模型学习进行正确的解耦计算，引入正交损失 L_{o} 作为约束：

$$L_{\text{o}} = \frac{1}{|B|}\sum_{i=1}^{|B|}\frac{\left|\left\langle \boldsymbol{t}_{\text{m}}^{(N+1)}(\boldsymbol{x}_i), \boldsymbol{t}_{\text{v}}^{(N+1)}(\boldsymbol{x}_i)\right\rangle\right|}{\left\|\boldsymbol{t}_{\text{m}}^{(N+1)}(\boldsymbol{x}_i)\right\|_2 \cdot \left\|\boldsymbol{t}_{\text{v}}^{(N+1)}(\boldsymbol{x}_i)\right\|_2} \tag{8}$$

式中：B 为一个批次的样本集合；$|B|$ 为获取批次数；$\boldsymbol{x}_i$ 为批次中的第 i 个样本。该损失函数强制最终输出的元令牌（身份特征）与视角令牌（视角特征）在向量空间尽可能正交，从而消除特征间的残留相关性。在监督信号上，全局令牌 $\boldsymbol{t}_{\text{m}}^{(N+1)}$ 接受行人身份标签的监督（结合交叉熵损失 L_{c}^{i} 与三元组损失 L_{t}^{i} 优化判别力），而视角令牌 $\boldsymbol{t}_{\text{v}}^{(N+1)}$ 则通过交叉熵损失 L_{c}^{v} 学习区分视角类别（如无人机/地面）。整体优化目标整合身份识别、视角分类与解耦约束，并与 1.2 节构建的属性三元组损失共同组成完整的目标函数：

$$L_{\text{total}} = L_{\text{attr}} + L_{\text{c}}^{\text{i}} + L_{\text{t}}^{\text{i}} + \lambda_L\left(L_{\text{c}}^{\text{v}} + L_{\text{o}}\right) \tag{9}$$

其中：λ_L 为超参数平衡任务权重；L_{attr} 为基于多模态属性解析的监督项。推理阶段仅使用精炼的元令牌 $\boldsymbol{t}_{\text{m}}^{(N+1)}$ 作为行人检索特征向量。

1.4 多尺度扩增 Transformer

多尺度扩增注意力机制模块由残差结构的深

度可分离卷积、归一化层、堆叠的扩增 Transformer、特征拼接层及最后的归一化与多层感知机(multi layer perceptron, MLP)层构成, 如图 2 所示。针对俯视角下行人特征分散的特性, 扩增 Transformer 在滑动窗口内执行差异化空洞采样: 对于位置 (i,j)的查询, 仅对以空洞率 r 的稀疏邻域进行键值提取(如 $(r=[1,2,3])$ 对应3×3至7×7感受野), 其坐标集合由 $(i'=i+pr, j'=j+qr)$ 定义 $(p,q\in[-1,1])$。模型利用了浅层注意力矩阵的局部性与稀疏性先验, 在避免全局计算冗余的同时, 通过通道分组的多头并行机制同步捕获行人局部纹理(小感受野)与肢体构型(大感受野)的跨尺度关联——不同头部分配独立空洞率后, 各组输出经拼接与线性投影融合为统一特征表达, 显著提升对高空俯视尺度压缩的鲁棒性。

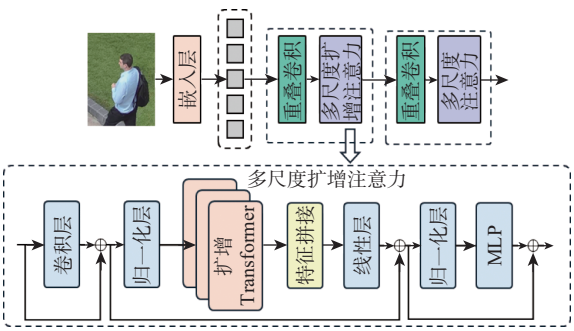


图 2　多尺度扩增注意力

Fig. 2　Multi-scale dilated attention

2　实验与结果分析

2.1　数据集与评估标准

本文在空地行人检索任务中采用两大主流基准进行性能验证:

AG-ReID.v1 [12] 数据集由一架在 15～45 m 高度飞行的无人机和大学校园内的地面摄像机拍摄, 包含 388 个行人身份的 21 983 张图像, 每个行人身份带有 15 个软属性。作为领域内首个小规模真实数据集, 重点评估跨视角泛化能力, 其核心协议包括空对地检索(A-G: 无人机→地面监控)和地对空检索(G-A: 地面监控→无人机)2 种定向模态变化场景。

AG-ReID.v2 [14] 包含了 1 615 个行人身份的 100 502 张图像, 由无人机、监控探头和可穿戴设备拍摄。具有地面检索(G-G)、空对空检索(A-A)、空对地检索(A-G)、空对穿戴设备检索(A-W)、地对空检索(G-A)、穿戴设备对空检索(W-A)6 种测试协议。本文采用其中的空对地检索(A-G)、地对空检索(G-A)协议进行实验。

CARGO [15] 作为大规模合成数据集, 包括 5 个地面摄像头、8 个空中摄像头、5 000 个身份及 108 563 张图像。CARGO 涵盖 4 类协议: 全视角混合检索(ALL)、纯地面域内匹配(G-G)、纯无人机域内匹配(A-A)及空-地跨视角泛化性测试(A-G)。

实验遵循 AG-ReID 通用评估框架, 联合采用三重指标全面量化模型性能: Rank-1 反映首候选命中率, 直接衡量跨视角检索精准性; 平均精度均值(mean average precision, mAP)综合评估召回率与查准率的平衡性; 平均逆负惩罚(mean inverse negative penalty, mINP)[22] 重点检验模型对相似干扰项的判别鲁棒性, 三者共同构建覆盖精度、鲁棒性与泛化能力的评测体系。

2.2　实验设置

本文方法在 NVIDIA GeForce RTX 4090 GPU 硬件平台基于 PyTorch 框架实现, 所有输入图像统一预处理为 256 像素×128 像素分辨率。本文方法构建了一行人结构化属性文件, 其中, 以 json 格式记录了多模态大模型解析的行人属性信息。训练模型采用 AdamW 优化器进行参数优化。针对 2 个数据集的训练超参数如下:

对于 AG-ReID.v1 数据集, 设置最大训练轮数为 80, 基准学习率为 0.01, 权重衰减系数为 0.000 5, 批次大小为 64, 预热策略采用初始比例因子为 0.1 配合 1 000 步渐进式学习率调整。

对于 AG-ReID.v2 数据集, 设置最大训练轮数为 200, 基准学习率为 0.001, 权重衰减系数为 0.000 5, 批次大小为 64, 并采用预热学习策略。

对于 CARGO 数据集, 设置最大训练轮数为 120, 基准学习率为 0.008, 权重衰减系数设置为 0.000 1, 批次大小为 128, 预热阶段采用比例因子为 0.01 并扩展至 2 000 步渐进调整, 以适应更大数据规模。

2.3　对比实验

本文方法在 AG-ReID.v1、AG-ReID.v2 和 CARGO 数据集上进行了测试。表 1 展示了本文方法在 AG-ReID.v1 数据集上的测试结果与现有方法的比较。实验结果表明, 本文方法在空对地检索和地对空检索中各指标上都达到了最优效果(见表中加黑数值)。相比 SD-ReID, 本文方法在空对地检索的 Rank-1、mAP、mINP 这 3 种指标上分别提升了 1.19%、0.42%、0.09%; 在地对空检索中分别提升了 0.35%、0.09%、0.31%。实验结果证明了本文方法在空地行人检索中身份有效判别了行人身份特征。

表 2 展示了本文方法在 AG-ReID.v2 数据集上与现有方法的性能比较, 加黑数值为最优结果。本文方法的 Ran-1 和 mAP 这 2 种指标在空对地(A-G)检索中较 V2E 方法提升了 0.15%、0.53%; 在地对空

表1 AG-ReID.v1数据集实验结果

Table 1 Experimental results on AG-ReID.v1 dataset

方法	Rank-1/%		mAP/%		mINP/%	
	A-G	G-A	A-G	G-A	A-G	G-A
OSNet[23]	72.59	74.22	58.32	60.99		
Bo[24]	70.01	71.20	55.47	58.83		
SBS[25]	73.54	73.70	59.77	62.27		
VV[26-27]	77.22	79.73	67.23	69.83	41.43	42.37
ViT[28]	81.28	82.64	72.38	73.35		
TransReID[7]	81.80	83.40	73.10	74.60		
FusionReID[29]	80.40	82.40	71.40	74.20		
CLIP-ReID[30]	79.44	84.20	70.55	73.05		
PCL-CLIP[31]	82.16	86.90	73.11	76.28		
Explain[13]	81.47	82.85	72.61	73.39		
VDT[15]	83.00	84.62	74.06	76.28	50.31	49.51
DTST[32]	83.48	84.72	74.51	76.05	49.86	50.04
SD-ReID[16]	85.16	85.97	75.40	77.02	51.35	50.42
本文方法	**86.35**	**86.32**	**75.82**	**77.11**	**51.44**	**50.73**

(G-A)检索中较V2E方法提升了0.17%、0.35%。实验结果表明,本文方法在身份较多、场景更多样的AG-ReID.v2数据集上也能取得良好效果,具有实际应用价值。

表2 AG-ReID.v2数据集实验结果

Table 2 Experimental results on AG-ReID.v2 dataset

方法	Rank-1/%		mAP/%	
	A-G	G-A	A-G	G-A
MGN[2]	82.09	84.21	70.17	72.41
BoT[24]	80.73	79.46	71.49	69.67
SBS[25]	81.96	84.10	72.04	73.89
V2E[14]	88.77	87.86	80.72	78.51
本文方法	**88.92**	**88.03**	**81.25**	**78.86**

表3展示了本文方法在CARGO数据集上的测试结果与现有方法的比较,加黑数值为最优结果。实验结果表明,本文方法在4种检索方式中各指标上都达到了最优效果。相比目前的最优方法SD-ReID,本文方法在空对地检索的Rank-1、mAP、mINP这3种指标上分别提升了1.45%、0.57%、0.49%;在混合检索场景中分别提升了1.19%、1.36%、1.30%;在地对地检索场景中分别提升了2.21%、0.96%、0.50%;在空对空场景中分别提升了0.23%、0.51%、0.34%。实验结果表明,本文方法通过解析行人图像属性有效地通过语义信息提升了模型身份判别能力,在不同检索场景中都提升了检索精度。

表3 CARGO数据集实验结果

Table 3 Experimental results on CARGO dataset

方法	Rank-1/%				mAP/%				mINP/%			
	A-G	ALL	G-G	A-A	A-G	ALL	G-G	A-A	A-G	ALL	G-G	A-A
SBS[25]	31.25	50.32	72.31	67.50	29.00	43.09	62.99	49.73	18.71	29.76	48.24	29.32
PCB[33]	34.40	51.00	74.10	55.00	30.40	44.50	67.60	44.60	20.10	32.20	55.10	27.00
BoT[24]	36.25	54.81	77.68	65.00	32.56	46.49	66.47	49.79	21.46	32.40	51.34	29.82
MGN[2]	31.87	54.81	83.93	65.00	33.47	49.08	71.05	52.96	24.64	36.52	55.20	36.78
VV[26-27]	31.25	45.83	72.31	67.50	29.00	38.84	62.99	49.73	18.71	39.57	48.24	29.32
AGW[22]	43.57	60.26	81.25	67.50	40.90	53.44	71.66	56.48	29.39	40.22	58.09	40.40
ViT[28]	43.13	61.54	82.14	80.00	40.11	53.54	71.34	64.47	28.20	39.62	57.55	47.07
VDT[15]	45.00	60.58	76.79	82.50	42.08	54.61	71.97	64.67	29.66	41.37	61.42	45.09
DTST[32]	50.63	64.42	78.57	80.00	43.39	55.73	72.40	63.31	29.46	41.92	62.10	44.67
SD-ReID[16]	53.12	65.06	81.25	82.50	46.44	57.47	74.08	67.70	33.03	44.21	63.10	51.90
本文方法	**54.57**	**66.25**	**83.46**	**82.73**	**47.01**	**58.83**	**75.04**	**68.21**	**33.52**	**45.51**	**63.60**	**52.24**

2.4 消融实验

为验证引入的3项改进措施的有效性,在AG-ReID.v1数据集上进行了消融实验,实验结果如表4所示,最优结果加粗显示。将TransReID作为基线模型,逐步加入属性解析模块、视角解耦架构,将Transformer层替换为多尺度扩增Transformer进行空地行人检索实验。实验结果表明,引入属性解析模块及属性三元组损失后,A-G检索任务Rank1、mAP、mINP分别提升7.48%、8.37%、5.19%,验证了属性监督对不同视角身份认定的强化作用;视角解耦架构叠加分层减法分离与正交损失模块,A-G检索3种指标较前阶段提升6.09%、5.23%、4.34%,改善视角差异导致的特征混淆;最后,整合多尺度扩增Transformer,模型在A-G检索中3种指标分别达到了86.35%、75.82%、51.44%,检索效果与前一阶段达到了相当水平,但参数量降低了11.7%,提升了计算速度降低了算力需求。消融实验表明,本文引入的3项改进对空地行人检索的精度及计算速度起到了优化作用。

表 4　消融实验

Table 4　Ablation experiment

方法	Rank-1/%		mAP/%		mINP/%		参数量
	A-G	G-A	A-G	G-A	A-G	G-A	
TransReID	72.63	78.36	62.61	65.37	41.59	42.30	8.050×10^{-7}
+属性	80.11	82.64	70.98	72.22	46.78	47.59	1.904×10^{-8}
+视角解耦	86.20	**86.83**	**76.21**	**77.38**	51.12	50.10	1.957×10^{-8}
+扩增Transformer	**86.35**	86.32	75.82	77.11	**51.44**	**50.73**	1.972×10^{-8}

为验证本文方法的效率优势与改进的实际部署价值, 在 AG-ReID.v1 数据集的测试实验中进行效率参数消融分析, 实验结果如表 5 所示。实验结果表明, 属性三元组损失由于加入了文本编码器, 因此, 模型参数量有比较大的提升, 推理速度有一定下降, 但仍然能保证模型推理的实时性。加入视角解耦架构改进微量增加了模型参数, 几乎未影响模型推理效率。将 Transformer 网络改进为扩增 Transformer 参数量下降至 1.729×10^{8}, 实现了参数的精简, 提升了模型的计算效率。

表 5　模型推理效率参数

Table 5　Model inference efficiency parameters

方法	帧率/(帧·s^{-1})	推理时间/s	参数量	浮点运算速度/10^{9} s^{-1}
TransReID	1 363	7.33×10^{-4}	8.050×10^{7}	11.13
+属性	758	1.32×10^{-3}	1.904×10^{8}	14.44
+视角解耦	749	1.33×10^{-3}	1.957×10^{8}	14.71
+扩增 Transformer	916	1.09×10^{-3}	1.729×10^{8}	12.04

2.5　可视化结果

为直观展示本文方法的行人检索效果, 在 AG-ReID.v1 数据集上进行了可视化实验, 展示了本文方法与 VDT 方法的检索效果对比及多模态大模型解析行人属性信息, 如图 3 所示。

如图 3(a)所示, 本文方法和 VDT 方法都正确检索出了相同身份的行人图像。VDT 方法检索出的匹配度前 5 的图像都是颜色风格相似的图像, 本文方法在成功检索颜色风格相似图像的基础上还检索出了阴影下的同一行人图像。结果表明, 本文方法的属性解析约束成功将颜色风格变化较大但行人属性相似的行人身份进行了正确关联。

如图 3(b)所示, 本文方法和 VDT 方法在本组检索结果中都出现了检索错误的行人图像。由于地面探头拍摄背景及行人形态与无人机拍摄的差异较大, VDT 方法匹配度前 4 的图像没有检索出身份正确的行人。本文方法由于进行了属性解析能够利用语义信息进行更准确的匹配并且引入多尺度扩增 Transformer, 能够更好地提取小像素图像(如匹配度第 3 的图像)的行人特征。然而, 由于无人机拍摄行人图像存在大量小像素行人图像, 多模态大模型对这类图像的属性解析效果较差, 对于具有相似属性的行人存在检索错误的情况。



图 3　空地行人检索可视化结果

Fig. 3　Visualization results of air-ground pedestrian retrieval

为检验视角解耦结构架构对特征分类的优化效果, 进行了 t-SNE 特征降维可视化与特征空间热力图可视化分析。

如图 4(a)～图 4(c) 所示, 红圈用于同一区域内 2 种方法的对比, 红圈内视角解耦架构的空地、地空行人检索方法的无人机与地面监控图像特征点相较基线方法距离更近, 表明本文方法特征空间中跨视角差异较小, 模型对于空间变化敏感度低。而基线架构的散点图红蓝散点倾向于分为两类, 如图 4(b)和图 4(d) 所示, 表明特征受不同视角影响较大。

图 5 展示了 AG-ReID.v1 测试的特征空间热力图结果。图 5(a) 为视角解耦方法测试的特征向量分布, 图 5(b) 为基线架构测试的特征向量分布。如

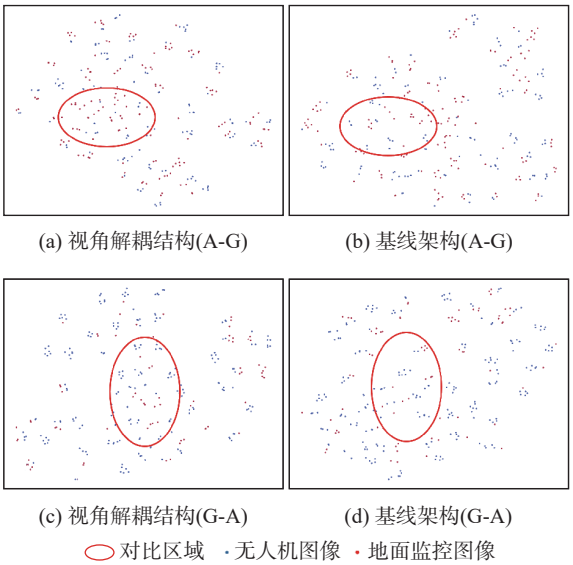


图 4 AG-ReID.v1 测试 t-SNE 可视化结果

Fig. 4 t-SNE visualization results of AG-ReID.v1 testing

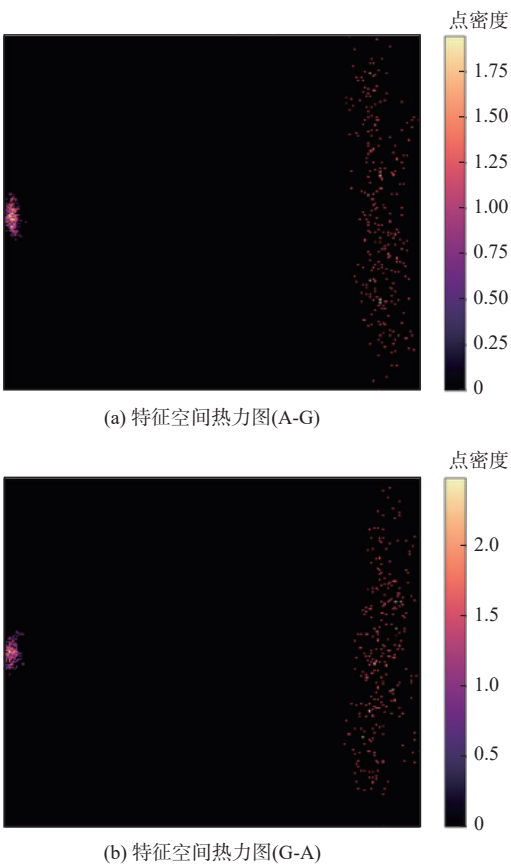


图 5 AG-ReID.v1 测试特征空间热力图

Fig. 5 Feature space heatmap of AG-ReID.v1testing

可视化结果所示，视角解耦方法的特征密度中心更重合，表明跨视角对齐更好；基线方法的特征分布较稀疏，表明跨视角表征稳定性较差。

3 局限性

尽管本文方法在构建基于属性解析的地空行人检索框架时，创新性地引入了多模态大模型的属性预测能力、视角解耦结构及扩增 Transformer 编码器，但仍面临一个普遍存在于跨视角行人检索特别是无人机场景的技术挑战：对低分辨率无人机图像进行可靠属性解析的固有局限性。无人机在高空远距离捕获的图像，其中包含行人的图像区域分辨率显著低于近处的地面监控视图，导致行人目标的像素信息稀疏、细小纹理体征高度模糊化。这种信息缺失会干扰属性解析模块的性能，特别是在预测精细或视觉区分度低的属性类别（如细粒度的衣着、鞋款、颜色）时，可能引入不可避免的预测偏差或不可置信度。尽管视角解耦机制旨在分离与视角无关的强鉴别特征，扩增 Transformer 在有限参数下捕捉更优的跨尺度特征表示，但这些架构上的改进仍然依赖于解析模块所提供的、源自目标像素的初步语义理解。因此，由低分辨率图像导致的属性解析置信度下降与潜在错误，可能在模型学习阶段成为偏差源，影响视角无关特征的表征力，并在最终的相似性度量中传导为检索错误。这突显了在低空成像条件下提升模型对低质信息鲁棒性的必要性，也是未来研究需重点攻克的难题之一。

4 结 论

本文针对无人机巡检场景下空地视角差异导致的行人检索性能瓶颈，提出一种属性语义引导的跨视角行人检索框架。主要结论如下：

1）多模态大模型解析空地行人属性，通过属性三元组损失有效约束网络匹配属性相似行人，显著提升俯视场景检索的精度。

2）引入分层减法分离与正交损失约束，进行解耦式特征建模，有效剥离视角相关干扰。

3）引入多尺度扩增 Transformer 优化 Transformer 网络，利用多尺度注意力机制提升小像素图像的表征能力，在保证模型准确度的同时降低了参数量。

实验验证表明，本文方法在 AG-ReID.v1 与 CARGO 数据集在空地、地空行人检索的 Rank1、mAP、mINP 指标上均优于当前主流模型，验证了该方法在无人机巡检场景下进行行人检索的有效性。

参考文献（References）

[1] Sun Y F, Zheng L, Yang Y, et al. Beyond part models: person retrieval with refined part pooling (and a strong convolutional baseline)[C]//Proceedings of the Computer Vision-ECCV 2018. Berlin: Springer, 2018: 501-518.

[2] Wang G S, Yuan Y F, Chen X, et al. Learning discriminative features with multiple granularities for person re-identification[C]//

Proceedings of the 26th ACM International Conference on Multimedia. New York: ACM, 2018: 274-282.

[3] Rao Y M, Chen G Y, Lu J W, et al. Counterfactual attention learning for fine-grained visual categorization and re-identification[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2022: 1005-1014.

[4] Zheng L, Shen L Y, Tian L, et al. Scalable person re-identification: a benchmark[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2016: 1116-1124.

[5] Wei L H, Zhang S L, Gao W, et al. Person transfer GAN to bridge domain gap for person re-identification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 79-88.

[6] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Berlin: Springer, 2017: 6000-6010.

[7] He S T, Luo H, Wang P C, et al. TransReID: Transformer-based object re-identification[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2022: 14993-15002.

[8] Zhang G W, Zhang P P, Qi J Q, et al. HAT: hierarchical aggregation Transformers for person re-identification[C]//Proceedings of the 29th ACM International Conference on Multimedia. New York: ACM, 2021: 516-525.

[9] Chen Y, Xia S X, Zhao J Q, et al. ResT-ReID: Transformer block-based residual learning for person re-identification[J]. Pattern Recognition Letters, 2022, 157: 90-96.

[10] Zhu H W, Ke W J, Li D, et al. Dual cross-attention learning for fine-grained visual categorization and object re-identification[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 4682-4692.

[11] Zhu K, Guo H Y, Zhang S L, et al. AAformer: auto-aligned transformer for person re-identification[J].IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(12): 17307-17317.

[12] Schumann A, Metzler J. Person re-identification across aerial and ground-based cameras by deep feature fusion[J]. Automatic Target Recognition XXVII, 2017, 10202: 102020A.

[13] Nguyen H, Nguyen K, Sridharan S, et al. Aerial-ground person re-ID[C]//Proceedings of the IEEE International Conference on Multimedia and Expo. Piscataway: IEEE Press, 2023: 2585-2590.

[14] Nguyen H, Nguyen K, Sridharan S, et al. AG-ReID. v2: bridging aerial and ground views for person re-identification[J]. IEEE Transactions on Information Forensics and Security, 2024, 19: 2896-2908.

[15] Zhang Q, Wang L, Patel V M, et al. View-decoupled transformer for person re-identification under aerial-ground camera network [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2024: 22000-22009.

[16] Wang Y H, Hu X, Wang L X, et al. SD-ReID: view-aware stable diffusion for aerial-ground person re-identification[EB/OL]. (2025-05-14)[2025-10-10]. https://arxiv.org/abs/2504.09549.

[17] Zheng A H, Pan P, Li H C, et al. Progressive attribute embedding for accurate cross-modality person re-ID[C]//Proceedings of the 30th ACM International Conference on Multimedia. New York: ACM, 2022: 4309-4317.

[18] Tan W T, Ding C X, Jiang J Y, et al. Harnessing the power of MLLMs for transferable text-to-image person ReID[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2024: 17127-17137.

[19] Yang S Y, Zhou Y N, Zheng Z D, et al. Towards unified text-based person retrieval: a large-scale multi-attribute and language search benchmark[C]//Proceedings of the 31st ACM International Conference on Multimedia. New York: ACM, 2023: 4492-4501.

[20] Jiao J Y, Tang Y M, Lin K Y, et al. DilateFormer: multi-scale dilated Transformer for visual recognition[J]. IEEE Transactions on Multimedia, 2023, 25: 8906-8919.

[21] Zeng A H, Xu B, Wang B W, et al. ChatGLM: a family of large language models from GLM-130B to GLM-4 all tools[EB/OL]. (2024-07-30)[2025-10-10]. https://arxiv.org/abs/2406.12793.

[22] Ye M, Shen J B, Lin G J, et al. Deep learning for person re-identification: a survey and outlook[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(6): 2872-2893.

[23] Zhou K Y, Yang Y X, Cavallaro A, et al. Learning generalisable omni-scale representations for person re-identification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 2021: 3069237.

[24] Luo H, Gu Y Z, Liao X Y, et al. Bag of tricks and a strong baseline for deep person re-identification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE Press, 2019: 1487-1495.

[25] He L X, Liao X Y, Liu W, et al. FastReID: a pytorch toolbox for general instance re-identification[C]// Proceedings of the 31st ACM International Conference on Multimedia. New York: ACM, 2023: 9664-9667.

[26] Kuma R, Weill E, Aghdasi F, et al. Vehicle re-identification: an efficient baseline using triplet embedding[C]// Proceedings of the International Joint Conference on Neural Networks. Piscataway: IEEE Press, 2019: 1-9.

[27] Kumar R, Weill E, Aghdasi F, et al. A strong and efficient baseline for vehicle re-identification using deep triplet embedding[J]. Journal of Artificial Intelligence and Soft Computing Research, 2020, 10(1): 27-45.

[28] Dosovitskiy A. An image is worth 16x16 words: Transformers for image recognition at scale[EB/OL]. (2021-06-03)[2025-10-10]. https://arxiv.org/abs/2010.11929.

[29] Wang Y H, Zhang P P, Liu X H, et al. Unity is strength: unifying convolutional and transformeral features for better person re-identification[J]. IEEE Transactions on Intelligent Transportation Systems, 2025, 26(3): 3713-3723.

[30] Li S Y, Sun L, Li Q L. CLIP-ReID: exploiting vision-language model for image re-identification without concrete text labels[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2023, 37(1): 1405-1413.

[31] Li J C, Gong X J. Prototypical contrastive learning-based CLIP fine-tuning for object re-identification[EB/OL]. (2025-01-14)[2025-10-10]. https://arxiv.org/abs/2310.17218.

[32] Wang Y H, Pishgar M. Dynamic token selection for aerial-ground person re-identification[EB/OL]. (2024-12-25)[2025-10-10]. https://

arxiv.org/abs/2412.00433.

[33] Sun Y F, Zheng L, Li Y L, et al. Learning part-based convolutional features for person re-identification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(3): 902-917.

Large model attribute parsing-based aerial-ground pedestrian retrieval method

TAN Quange[1], WANG Rong[1], LIAO Mancheng[2], LI Xin[1,*]

(1. College of Information Network Security, People's Public Security University of China, Beijing 100038, China;
2. College of Criminal Investigation, People's Public Security University of China, Beijing 100038, China)

Abstract: Synergistic analysis of drone-captured imagery and fixed surveillance video enables continuous tracking of targets in inspection tasks, achieving cross-view person re-identification in areas where surveillance probes are sparsely distributed. The transferability of image retrieval algorithms created for conventional surveillance situations to aerial-ground cross-view environments is limited by the notable discrepancy between the horizontal view of ground surveillance and the bird's-eye view of drones. Existing aerial-ground pedestrian retrieval methods primarily focus on mitigating the appearance discrepancies caused by cross-view variations, while the mining and analysis of person attribute characteristics remain insufficiently explored. To address these issues, this paper proposes a aerial-ground pedestrian retrieval method based on large model attribute parsing. A person parsing module is constructed based on a multi-modal large model to generate fine-grained semantic attributes, and an attribute triplet loss is designed to achieve cross-view semantic alignment. A view decoupling architecture is introduced to separate view-specific features through hierarchical subtraction, with an orthogonal loss applied to constrain feature independence. A multi-scale dilated Transformer is incorporated, combining multi-scale dilated attention with global self-attention to optimize the balance between computational complexity and receptive field, thereby reducing model parameters. The effectiveness of the methodology is confirmed by experiments on the AG-ReID.v1, AG-ReID.v2, and CARGO datasets, which show that the suggested strategy successfully increases performance on Rank-1, mAP, and mINP metrics in aerial-ground pedestrian retrieval tasks.

Keywords: UAV; pedestrian retrieval; aerial-ground imagery; attribute parsing; multimodal large language models

Received: 2025-12-02; **Accepted**: 2026-01-25; **Published Online**: 2026-04-03 10:45

URL: link.cnki.net/urlid/11.2625.V.20260402.1729.004

Foundation item: National Natural Science Foundation of China (62076246)

* **Corresponding author**. E-mail: lixin@ppsuc.edu.cn