中图法分类号： **文献标识码：** A **文章编号：** 1006-8961(XXXX)XX-0001-16

论文引用格式： Zhou Xinyu, Wu Yu, Cao Min, Zhu Qiaoming, Ye Mang. XXXX. A benchmark dataset for multi-view mixed text-to-image person retrieval. Journal of Image and Graphics, XX(XX):0001-0016(周新宇, 吴彧, 曹敏, 朱巧明, 叶茫. XXXX. 文本-图像多视角混合行人检索基准数据集研究. 中国图象图形学报, XX(XX):0001-0016)[DOI:10.11834/jig.250604]

文本-图像多视角混合行人检索基准数据集研究

周新宇[1],吴彧[1],曹敏[1*],朱巧明[1],叶茫[2]

1. 苏州大学计算机科学与技术学院,苏州 215008; 2. 武汉大学计算机学院,武汉 430072

摘 要：目的 文本-图像的行人检索任务(text-to-image person retrieval, TIPR)利用自然语言描述从监控图像库中定位目标行人,是智慧城市与智能安防中的关键技术。随着无人机成像技术的发展,“空-地协同”多视角监控成为主流趋势,但目前TIPR研究主要聚焦地面视角,缺乏同时涵盖空中与地面视角的基准数据集,制约了其在多视角协同监控场景中的应用。因此,本文旨在构建一个包含空中与地面视角的混合多视角行人检索基准数据集,以推动多视角TIPR的研究。**方法** 为此,本文构建了文本-图像的空中-地面多视角混合行人检索数据集(aerial-ground mixed-view person retrieval dataset, AGMix-PR)。该数据集通过采集多源数据库中的空中与地面视角行人图像,真实反映了实际应用场景中的挑战。为支持大规模高质量文本标注,本文进一步提出一种属性引导的层次化文本生成框架:利用结构化属性作为高层引导信号驱动多模态大语言模型生成准确且多样化的自然语言描述,确保了低成本条件下的高质量文本输出。**结果** 实验结果表明,现有主流TIPR方法在AGMix-PR上整体性能显著下降,反映出多视角场景带来的挑战。**结论** AGMix-PR数据集有效弥补了当前TIPR领域在空-地多视角混合场景下基准数据集的不足,其构建方法为经济高效地创建多模态数据集提供了新思路。该数据集为相关研究提供了一个重要的基准平台,有助于推动鲁棒行人检索理论与方法的发展,适应真实复杂场景的需求。AGMix-PR数据集的在线发布地址为:https://github.com/barry-wy/AGMix_PR。

关键词： 文本-图像检索;文本-图像的行人检索;行人重识别;多模态学习;数据集

A benchmark dataset for multi-view mixed text-to-image person retrieval

Zhou Xinyu[1], Wu Yu[1], Cao Min[1*], Zhu Qiaoming[1], Ye Mang[2]

1. School of Computer Science and Technology, Soochow University, Soochow 215008, China; 2. School of Computer Science, Wuhan University, Wuhan 430072, China

Abstract: Objective Text-to-image person retrieval (TIPR) aims to identify and retrieve the target pedestrian from large-scale surveillance image galleries based on natural language descriptions. Compared with the structured textual queries commonly used in person re-identification (Re-ID), natural language descriptions are more flexible, open-ended, and easier to obtain. As a result, TIPR has become a crucial component in intelligent public security systems, smart city infrastructures, and large-scale urban surveillance platforms. Recently, the rapid expansion of the low-altitude economy, together with the widespread deployment of unmanned aerial vehicles (UAVs) for urban sensing, traffic monitoring, emergency response, and public safety, has given rise to hybrid aerial-ground surveillance systems. In such systems, pedestrians may be observed simultaneously or alternately from aerial platforms and ground cameras, resulting in drastic variations

收稿日期: 2025-11-30;**修回日期:** 2026-01-19

* **通信作者:** 曹敏 caomin0719@126.com

基金项目: 国家自然科学基金项目(62476188)

Supported by: National Natural Science Foundation of China (62476188)

in scale, camera elevation, viewing angle, occlusion patterns, illumination, and background context. Although substantial progress has been made in TIPR, existing TIPR studies are developed and evaluated under a highly constrained assumption: pedestrians are captured exclusively by fixed, ground-based surveillance cameras with relatively stable viewpoints and consistent imaging conditions. Benchmark datasets that simultaneously encompass both aerial and ground viewpoints are still scarce. Moreover, the few available datasets often suffer from limited scale and oversimplified annotation protocols, making them insufficient for modeling the complexity of real hybrid-view environments. As a result, current TIPR methods often generalize poorly when deployed in aerial-ground scenarios. **Method** To bridge this gap, we introduce AGMix-PR (aerial-ground mixed-view person retrieval dataset), a novel benchmark dataset specifically designed for text-to-image person retrieval in aerial-ground multi-view mixed scenarios. AGMix-PR integrates pedestrian data collected from both UAV-mounted cameras and fixed ground surveillance systems across diverse urban and suburban environments. The dataset comprises over 8,000 unique pedestrian identities and more than 70,000 image-text pairs, making it one of the largest and most comprehensive TIPR datasets to date that explicitly bridges aerial and ground viewpoints. Each identity is associated with images captured under heterogeneous conditions, including extreme scale variations, drastic viewpoint changes, partial occlusions, and cluttered backgrounds, thereby reflecting the intrinsic challenges of real-world hybrid surveillance systems. A key challenge in constructing AGMix-PR lies in obtaining high-quality, fine-grained textual descriptions at scale without prohibitive human annotation costs. To this end, we propose an *attribute-guided hierarchical text generation framework* that decouples the complex language generation task into two synergistic stages. In the low-level attribute prediction stage, we leverage the pre-trained UniHCP model, a state-of-the-art attribute classifier, to automatically extract a structured vector of pre-defined semantic attributes per image (*e. g.*, gender, age, upper or lower clothing type and color, footwear, accessories). In the high-level text generation stage, these structured attributes are used as controllable priors to guide a multimodal large language model (MLLM) to generate high-quality, controllable textual descriptions. We design two complementary prompting strategies: (1) attribute-guided generation strategy, where randomized attribute sequences are embedded into a directive instruction to encourage syntactic diversity; and (2) template-guided generation strategy, where some handcrafted linguistic templates abstracted from real-world descriptions, are randomly selected to enforce coherent narrative structure. Both strategies explicitly suppress background mentions and hallucinated details. The final output for each image consists of two distinct yet semantically consistent captions, significantly enriching linguistic variation while preserving factual accuracy. Meanwhile, a rigorous text-cleaning pipeline is adopted to assess and enhance the quality of generated textual descriptions. This pipeline contains automatic CLIPScore-based filtering, through which texts are categorized into clean and noisy samples. For the generated texts which may contains noise, we employ a differentiated handling strategy tailored to the dataset split. Specifically, we deliberately retain noisy samples in the training set to preserve the large-scale nature of the dataset and to reduce the risk of overfitting to an unrealistically clean distribution. In contrast, all noisy samples detected in the test set are carefully corrected by expert human annotators, thus providing a fair and reliable evaluation. **Result** We conduct extensive experiments on AGMix-PR using a diverse set of representative state-of-the-art TIPR methods. Experimental results consistently demonstrate that, compared to conventional ground-only settings, existing TIPR models suffer a notable performance drop when evaluated under aerial-ground mixed-view conditions. This observation highlights the severity of the cross-view gap and exposes the limitations of current methods in handling extreme viewpoint variations. More importantly, further analyses reveal that models trained on the full AGMix-PR dataset—leveraging both aerial and ground images—consistently outperform those trained on single-view subsets across all evaluation protocols. These findings validate our core hypothesis that multi-view supervision is essential for learning view-invariant, semantically grounded person representations in TIPR. **Conclusion** The AGMix-PR benchmark establishes a new standard for evaluating TIPR methods in realistic, aerial-ground mixed view surveillance environments. By bridging the aerial and ground views, it not only exposes the limitations of current approaches but also provides a fertile ground for developing view-robust, semantically grounded retrieval models. The proposed attribute-guided hierarchical generation pipeline offers a scalable, cost-effective paradigm for constructing large-scale multimodal datasets, demonstrating that structured intermediate representations can dramatically enhance the quality and controllability of synthetic annotations. The dataset is available athttps://github. com/barry-wy/AGMix_PR.

Key words: text-image retrieval; text-to-image person retrieval; person re-identification; multi-modal learning; dataset

0 引 言

文本-图像的行人检索(text-to-image person retrieval, TIPR)是计算机视觉与自然语言处理相结合的跨模态任务,其核心目标是以自然语言描述作为查询条件,从大规模行人图像库中精准地检索出与该描述相匹配的目标人物图像。相较于传统行人重识别(person re-identification, Re-ID)依赖图像进行检索的方式, TIPR的关键优势在于用更灵活、易获取的文本描述取代查询图像,显著提升了检索系统的实用性。这使其尤其适用于仅有目击者文字描述而无法获取目标图像的现实场景,在嫌疑人员追踪、智能安防与视频监控等方面有重要应用价值。因此, TIPR近年来受到广泛的研究(Bai等,2023b; Song等,2024; Tan等,2024; Cao等,2025)。

然而,当前主流研究(Qin等,2024; Jiang和Ye, 2023; Jiang等,2025)普遍基于传统的地面监控体系,默认行人样本均由固定位置的地面摄像头采集。这一设定在过于简化了现实场景中复杂多样的成像特点。随着低空经济在2025年《政府工作报告》中被列为战略性新兴产业,无人机成像技术快速发展(肖云等,2025; 王旭辰等,2022; 冷佳旭等,2023),推动了"空-地协同"多视角监控架构的形成。传统地面摄像头作为静态设备,受限于固定安装位置与视野遮挡,难以有效覆盖广场、交通枢纽及大型活动现场等开阔区域;而无人机作为可移动的动态空中平台,不仅能够通过空中俯视或斜向观测,突破地理障碍,拓展监控范围,有效弥补地面视角的盲区,还能对目标行人实施灵活、持续的追踪。二者的协同部署既显著提升了监控系统在复杂开放场景下的多视角立体感知能力,又有效融合了静态广域覆盖与动态精准追踪的优势,已成为Re-ID领域的研究热点之一(Zhang等,2023; Nguyen等,2024; Wang等,2025b)。然而, TIPR领域在该多视角混合场景下的研究仍较为稀缺,相关基准数据集(Zhou等,2025; Wang等,2025b)也较为有限,并普遍受限于数据集规模以及较为简单的标注文本。

针对上述不足,本文创新性地提出一个文本-图像的空中-地面混合视角行人检索数据集(aerial-ground mixed-view person retrieval dataset, AGMix-PR)。该数据集在规模、视角覆盖与标注体系上均实现全面扩展,不仅包含更大规模的多视角行人图像,还补充了细粒度的属性标注以增强数据的可用性与信息密度。相较于通用图像描述任务(Vinyals等,2016; Li和Chen,2018; Zhang等,2021; Xu等,2025)仅需对图像中的主要对象及整体语义内容进行概括性描述,对行人的文本标注具有更高的专业性与细粒度要求。如图1所示,行人服饰的类别与颜色、携带物品等属性需要被精细描述。因此,该标注对标注人员的专业素养要求较高,

图1 通用领域与行人领域文本描述图

Fig. 1 Text description between general domain and person domain ((a)text description of general domain(coarse-grained); (b) text description of person domain(fine-grained))

人力成本高。同时,混合视角数据普遍存在的低分辨率、视角变换及遮挡等问题,进一步提升了标注难度。随着多模态大模型(Bai等,2023a; Li等,2024; Wang等,2024)的快速发展,其在图像语义理解与自然语言生成方面展现出强大能力,为自动化生成高质量、细粒度行人描述提供了新路径。在此基础上,本文构建了一种属性引导的层次化文本生成框架。该框架首先通过相对简单的行人属性标注任务实现对图像内容的细粒度感知;随后将所得结构化属性作为语义先验,引导多模态大语言模型生成语义准确且句式多样的文本描述,有效降低了生成难度,能够实现高效地获取高质量文本标注。此外,混合视角数据在成像几何、尺度变化、背景复杂度等方面存在显著差异,导致不同视角特征分布存在一定偏移,使得传统面向地面视角设计的方法在空-地场景下面临性能退化问题。因此,本文在

AGMix-PR 数据集上系统评估现有主流 TIPR 方法，揭示他们在混合视角条件下的局限性，为后续方法创新提供研究基础。

本文主要贡献如下：1）构建了一个文本-图像的空中-地面混合视角行人检索数据集。该数据集更贴近真实复杂监控场景中多源异构视角共存的场景。2）提出了属性引导的层次化文本生成框架，在显著降低人工干预的同时实现低成本、高质量的自动化行人文本描述生成。3）在 AGMix-PR 上对多种代表性的TIPR 方法进行了全面实验评估，深入分析了现有技术在文本-图像的空中-地面行人检索任务的瓶颈与挑战，为未来研究提供基准支撑。

1 相关工作

1.1 文本–图像的行人检索数据集

CUHK-PEDES（Li 等，2017a）是首个面向文本-图像的行人检索的基准数据集。该数据集包含 13 003 个行人身份的 40 206 张图像和 80 412 条文本描述，是该领域最具代表性的基准数据集之一。然而，CUHK-PEDES 中的文本包含大量与行人外观无关的细节信息，例如人物动作或者背景。为解决这一问题，Ding 等人（2021）构建了 ICFG-PEDES 基准数据集，其文本以行人为中心，着重描述行人的外观特征。与此同时，RSTPReid 数据集（Zhu 等，2021）则致力于更真实地反映实际监控场景的复杂性。该数据集覆盖 15 个相机、室内/室外与不同时段，反映了同一摄像网络下的真实监控场景。此外，为进一步提升文本描述的粒度，Zuo 等人（2024）构建了 UFine-Bench 基准数据集，其核心子集 UFine6926 中文本描述的平均长度达到约 80.8 词，达到其他主流数据集的 2-3 倍。表 1 详细列举了主流的行人检索数据集。

尽管上述数据集在推动文本-图像的行人检索任务发展方面发挥了积极作用，但它们均局限于地面摄像头视角。Wang 等人（2025b）首次提出了面向空-地的 TBAPR 数据集，其设定要求每个行人的图像集同时包含来自空中和地面的成对视角，为多视角场景下的行人检索研究提供了新的基准。然而，这一设定在实际监控系统中较为理想化：现实场景中难以保证所有行人均具备同时覆盖空-地的成对图像，因此该数据集的场景适用性存在一定限制。随后提出的 TAG-PEDES（Zhou 等，2025；Wang 等，2025b）放宽了这一约束，允许不同行人身份有多样化的视角组合，更加贴合真实监控系统中多源异构成像的特点。然而，该数据集在文本标注方面仍依赖相对朴素的生成策略，缺乏对细粒度属性信息的标注和利用，同时整体数据规模也较为有限。基于上述不足，本文在这些工作的基础上进一步扩展数据规模，并以行人属性为核心锚点构建属性引导的层次化文本生成框架，从而显著提升标注文本的可控性、细粒度性与语义准确度，为空－地混合场景下的TIPR 研究提供更加全面、可靠的基准支撑。

1.2 行人重识别的数据集

行人重识别旨在以一张特定行人的查询图像作为输入，从大规模图像库中检索出属于同一身份的其他行人图像。该任务早期主要聚焦于地面监控场景，相关数据集也普遍基于固定地面摄像头采集。例如，Market-1501 是由 Zheng 等人（2015）提出的地面视角的行人重识别数据集，包含 1 501 个行人身份的 32 668 张图像。近年来，随着无人机技术与低空遥感平台的快速发展，行人重识别研究正逐步拓展至空中视角，催生了面向无人机拍摄场景的空中行人重识别任务。PRAI-1581（Zhang 等，2020）作为首个专注于空中行人重识别的基准数据集，包含 1 581 个行人的 39 461 张图像，为该领域的研究奠定了重要基础。随后，UavHuman 数据集（Li 等，2021b）的提出在规模、多样性和现实挑战性方面实现了重要突破。该数据集包含 1 144 名行人在不同环境条件下由无人机低空拍摄的 41 290 张图像，涵盖多样的光照、姿态与复杂背景，全面模拟了真实航拍环境下的典型困难，显著提升了空中行人检索任务的实用价值与研究深度。

在此基础上，Nguyen 等人（2023）首次提出了 AG-ReID. v1 数据集以推动空中-地面行人重识别研究。该数据集包含同一行人在空中与地面不同视角下的配对图像，为该任务提供了初步的研究基准。随后的研究工作（Nguyen 等，2024）持续扩展该数据集的规模并引入更多样的视角变化，进一步提升了任务的复杂性与实用性。此外，Zhang 等人（2024）利用虚拟引擎获取多视角行人图像，构建了首个大规模合成数据集 CARGO，有效缓解了真实数据采集与标注的高成本问题。然而，这些现有的数据集均局限在单一的视觉模态，缺少跨模态语义信息。相

表 1　行人检索数据集对比表

Table 1　Comparison of person retrieval dataset

任务	模态	数据集名称	身份数	图片数量	文本数量	图像来源	文本来源	视角类型	提出年份
行人重识别	图像	Market-1501	1 501	32 668	-	真实收集	-	地面	2015
		PRAI-1581	1 581	39 461	-	真实收集	-	空中	2020
		UavHuman	1 144	41 290	-	真实收集	-	空中	2021
		AG-ReID.v1	388	21 983	-	真实收集	-	空中&地面	2023
		AG-ReID.v2	1 615	100 502	-	真实收集	-	空中&地面	2024
		CARGO	5 000	108 563	-	模拟合成	-	空中&地面	2024
文本-图像的行人检索	文本&图像	CUHK-PEDES	13 003	40 206	80 440	真实收集	人工标注	地面	2017
		ICFG-PEDES	4 102	54 522	54 522	真实收集	人工标注	地面	2021
		RSTPReid	4 101	20 505	41 010	真实收集	人工标注	地面	2021
		UFine6926	6 926	26 206	52 412	真实收集	人工标注	地面	2024
		TAG-PEDES	6 840	28 178	56 356	真实收集	模型生成	空中&地面	2025
		TBAPR	2 238	65 880	65 880	真实收集	模型生成	空中&地面	2025
		AGMix-PR	8 363	37 344	74 688	真实收集	模型生成	空中&地面	2025

注："-"表示具体信息未知或不存在。

比之下，本文提出的 AGMix-PR 数据集不仅包含空中与地面视角的图像，还利用自动化标注提供了丰富的文本描述，为文本-图像的空中-地面行人检索提供了新的基准与研究方向。

1.3　文本-图像的行人检索方法

近年来，文本-图像的行人检索方法迅速发展。早期研究中，Li 等人（2017b）和 Chen 等人（2021）主要采用独立训练的单模态编码器，例如基于经典卷积神经网络（Simonyan 和 Zisserman，2014）的图像编码器和基于长短期记忆递归神经网络（long short-term memory，LSTM）（Graves，2012）的文本编码器，分别提取图像与文本的全局特征以实现跨模态对齐，然而该方法难以建立细粒度的语义关联。为克服这一局限，后续研究 Jing 等人（2020）引入显式的对齐机制，借助外部工具实现图像局部区域与文本词汇间的细粒度对齐。最近，视觉-语言预训练模型（Radford 等，2021；Li 等，2021a）因其强大的跨模态表示能力而引起了极大的关注。基于此类模型，多项研究（Jiang 和 Ye，2023；Cao 等，2024；Yan 等，2024；Yang 等，2023）在 TIPR 任务上实现了显著性能提升。例如，Cao 等人（2024）对对比语言-图像预训练模型（contrastive language-image pre- training，CLIP）（Radford 等，2021）模型在细粒度对齐能力方面进行了系统评估。然而，这些方法主要聚焦于地面监控视角下的行人检索方法，鉴于在实际多源监控场景中，无人机拍摄的空中视角图像与地面摄像头获取的图像在视觉特征上存在显著差异：无人机图像多为俯视或倾斜俯视视角，常出现头部主导、身体遮挡或多角度形变等情况。因此，现有文本-图像的行人检索方法，亟需在空中-地面视角混合的行人检索数据集上，进行全面评测验证，以揭示其在多视角条件下的适用性与潜在局限性。

2　数据集

本章将从数据集的构建方法与统计特性两方面对提出的 AGMix-PR 数据集进行详细介绍。

2.1　数据集构建

本节详细介绍 AGMix-PR 数据集的构建过程，包括三个关键环节：多视角图像收集、属性引导的层次化文本生成，以及系统化数据清洗。

1）　多视角图像收集

为更全面地反映真实世界场景中行人视角的多样性与复杂性，本文整合多个现有行人数据集，构建

(a) 地面视角示例图

(b) 空中视角示例图

(c) 空-地视角示例图

((a)examples of ground view; (b)examples of aerial view; (c) examples of aerial-ground view)

图2 不同视角下行人图像示例

Fig. 2 Example images of pedestrians from different viewpoints

了一个涵盖空中-地面多视角的图像集合。具体而言,该集合涵盖以下三类来源:(1)地面视角图像,源自数据集 Market-1501(Zheng等,2015);(2)空中视角图像,来自无人机视角数据集:PRAI-1581(Zhang等,2020)和 UavHuman(Li等,2021b);(3)空中-地面多视角图像,源于两个同时包含空中-地面视角的行人重识别数据集:AG-ReID. v2(Nguyen et al., 2024)、LAGPeR(Wang et al., 2025)和 G2APS(Zhang et al., 2023)。如图 2 所示,不

表 2 AGMix-PR数据集图像库构成

Table2 Composition of the image gallery in AGMix-PR

视角类型	图片来源	身份数	图片数量
地面	Market-1501	1 521	7 104
空中	PRAI-1581	842	3 085
	UavHuman	633	2 241
空中&地面	AG-Reid.v2	1 599	7 049
	G2APS	2 534	9 530
	LAGPeR	1 521	7 104

同视角下的行人图像存在显著差异。传统地面摄像头通常以较低视角从正面或侧面拍摄,能够较完整地呈现行人的细节特征;而无人机获取的空中视角图像则多为自上而下的俯视角度,并常伴随大幅度的旋转与尺度变化。因此,即使是同一行人,不同的视角会带来一定外观上的变化,反应了真实检索场景下的难点。

在收集完所有图像后,经过一轮初步的筛查,其中的低质量数据被剔除了,包括分辨率过低(图像像素总数 ≤ 1 000)的样本,以及仅包含部分人体的不完整图像,整体数据质量与多样性均得到有效保障。最终构建了一个包含 32 243 张图像,涵盖 8 229 个不同的行人身份的图像库。表 2 详细列出了本数据集图像库的构成情况。

2) 属性引导的层次化文本生成

视角的多样性在提升数据集真实性的同时,也引入了图像分辨率偏低、视角变化剧烈、遮挡更为严重等问题,显著增加了高质量文本描述生成的难度。为此,本文提出一种属性引导的层次化文本生成框架:首先对行人图像进行视觉属性感知,提取服饰类别、颜色、配饰等细粒度特征;进而将这些结构化属性作为先验知识,引导多模态大模型生成准确、一致的自然语言描述。该框架通过低层属性标注到高层文本生成的级联式设计,能够有效提升自动化文本标注的质量与可靠性。

(1)属性标注。为获取高质量的结构化属性标注,本文对 AGMix-PR 中的所有行人图像统一采用基于预训练模型的自动化属性标注策略。首先,本文预定义了一个囊括关键行人属性的集合 $\boldsymbol{A}$,$\boldsymbol{A}=\{a_k \mid k \in [1, N_a]\}$,式中,$N_a$ 为预定义属性数量,元素 a_k 代表对应了一个具体的属性,例如性别、是否有眼镜等等,如表 3 所示。对于任意一张行人图像 $\boldsymbol{I}$,预训练行人属性识别模型(unified model for human-centric perceptions, UniHCP)(Ci等,2023)被用于预测其属性置信度向量 $\boldsymbol{p}$:

$$\boldsymbol{p} = UniHCP(\boldsymbol{I}) \tag{1}$$

式中,$\boldsymbol{I}$ 代表输入的行人图像,$UniHCP$ 代表经过预训练后的行人属性模型,$p \in [0,1]^{N_a}$ 为输出的属性置信度向量,其第 k 维($k = 1, 2, \ldots, N_a$)表示第 k 个属性存在的概率,即为 p^k。随后对初步的置信度 $\boldsymbol{p}$ 进行了三阶段的后处理:①置信度过滤。设定阈值 $\theta = 0.7$,仅 $p^k > \theta$ 的高置信度属性会被置为 1,其余属性被置为 0。该阈值的选择基于对数据集属性概率分布的统计分析:绝大多数属性(如 Female、Short-Sleeve、Trousers 等)的概率分布呈现明显的双峰特性,高密度区域集中在接近 0 和接近 1 的两端,而区

表 3　行人属性标注定义

Table 3　Definition of Annotated Pedestrian Attributes

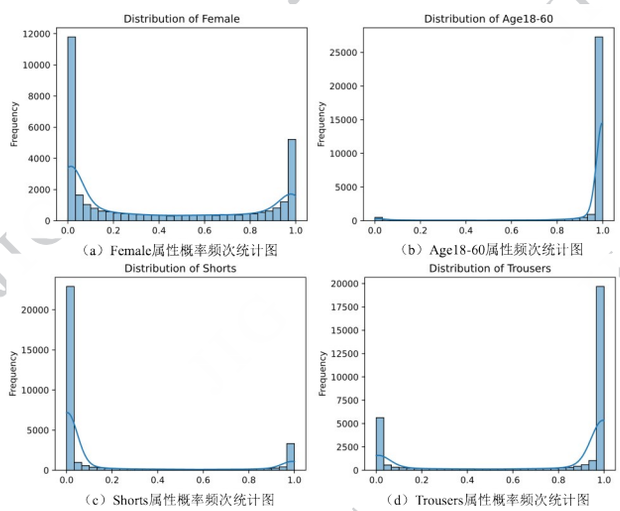
属性类别	属性名	标注
性别	Gender	女(1)男(0)
年龄	AgeOver60, Age18-60, AgeLess18	年龄大于60 [0/1]、在18-60之间 [0/1]、小于18 [0/1]
袖子长度	Short, Long	短袖 [0/1]、长袖[0/1]
眼镜	Glasses	有(1)无(0)
帽子	Hat	有(1)无(0)
靴子	Boots	是(1)否(0)
手持物品	HoldObjectsInFront	是(1)否(0)
上衣款式	LongCoat	长款大衣[0/1]
下身款式	Trousers, Shorts, Skirt&Dress	长裤[0/1]、短裤[0/1]、裙子/连衣裙[0/1]
上衣图案	Stride, Logo, Plaid, Splice	有条纹[0/1]、有Logo[0/1]、格子图案[0/1]、结构拼接[0/1]
下身图案	Stripe, Pattern	下身有条纹[0/1]、下身有图案[0/1]
包的款式	HandBag, ShoulderBag, Backpack	手提包[0/1]、单肩包[0/1]、背包[0/1]
人体朝向	Front, Side, Back	人体朝前[0/1]、朝侧[0/1]、朝后[0/1]

注："[0/1]"表示"是"为"1"，"否"为"0"。

间（0.2，0.7）内的样本占比极低，表明模型对属性判别具有高度置信度，如图3所示。选择 $\theta = 0.7$ 可有效过滤低置信度预测，同时保留高可靠性正样本，避免因模糊边界引入噪声。

②冲突属性消解。对于相互冲突的属性，例如同一属性类别中多个标注（如AgeOver60、Age18-60和AgeLess18），置信度最高的属性标签会被保留。

③标准化短语映射。将获得的独热编码形式的标签映射会对应的属性名，获得自然语言形式的属性标签。这些属性标签不仅能够作为细粒度的标注信息辅助模型学习，而且可以作为后续文本生成过程中的关键先验知识。



（a）Female属性概率频次统计图　（b）Age18-60属性概率频次统计图

（c）Shorts属性概率频次统计图　（d）Trousers属性概率频次统计图

图 3　数据集属性概率分布图

Fig. 3　Distribution diagram of attribute probability ((a) frequency statistics chart of Female attribute probabilities; (b) frequency statistics chart of Age18-60 attribute probabilities; (c) frequency statistics chart of Shorts attribute probabilities; (d) frequency statistics chart of Trouser attribute probabilities)

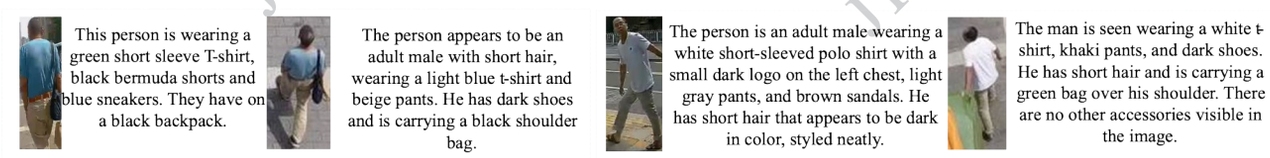
（2）文本生成。在获得结构化属性的基础上，本文利用多模态大语言模型构建了一种兼顾语义准确性、语言多样性与风格可控性的文本生成范式。该范式融合两种互补的生成方式：属性引导生成以及模板引导生成。典型样例，如图4所示。

①属性引导生成，通过将结构化属性直接嵌入提示词，实现对生成内容的细粒度语义控制。具体而言，对于每张行人图像，一个固定的提示词模板被用于驱动多模态大语言模型生成文本：*Don't mention the background of the people in the image. Please provide a detailed description of this person's <attributes>. Finally, combine all the details into a single sentence.* 其中的占位符<*attributes*>会被替换为该行人对应的属性，例如female，age 18 - 60等等。同时，为避免因固定属性顺序导致的句式僵化，属性序列

中会引入随机的顺序扰动，使模型在严格保留所有关键属性的前提下，灵活调整主谓宾结构、修饰语位置及连接逻辑，从而生成语法多样且语义一致的自然语言描述。

(a) 地面视角图文对示例图

(b) 空中视角图文对示例图



(c) 空-地视角图文对示例图

图 4　AGMix-PR数据集图文对示例图

Fig. 4　Examples of image-text pair in AGMix-PR ((a)examples of image-text pair of ground view; (b)examples of image-text pair of aerial view; (c)examples of image-text pair of aerial-ground view)

②模板引导生成，则进一步引入结构化语言先验，以提升文本的规范性与可读性。受现有对行人描述文本可控生成研究(Tan等，2024)的启发，本文基于AGMix-PR的属性体系，利用大语言模型(Zhu等，2023)离线生成并人工筛选出多个高质量描述模板，覆盖常见行人外观描述模式，例如：*The <geneder> is dressed in <clothing attributes>, <footwear attributes>, <accessory attributes>, and carrying <bag attributes>.* 对于这些占位符中的已知属性会被替换。在生成阶段，系统随机选取一个模板作为风格参考，并构造如下增强型提示词：*Generate a description about the overall appearance of the person, in a style similar to the template: <template>. If some requirements in the template are not visible, you can ignore.* 其中占位符<*template*>随机替换为预先选取的文本模板。这一策略使模型在多种语言风格与描述粒度下保持正确性。

最终，每张行人图像都通过属性引导生成和模板引导生成两种方式获得两条高质量文本标注，在保障语义准确性的同时显著增强了文本表达多样性。

上述文本生成过程基于多模态大语言模型LLaVA-OneVision(large language and vision assistant onevision)(Li等，2024)实现。具体而言，该模型采用Qwen2(Team，2024)作为语言主干模型，具备强大的语言理解与生成能力；视觉感知模块选用SigLIP (sigmoid loss for language image pre-training)(Zhai等，2023)模型，以384×384分辨率输入图像，通过图像块尺寸为14的视觉Transformer(vision transformer, ViT)(Dosovitskiy，2020)架构提取全局视觉特征，并经由可学习的投影层与语言模型对齐。在推理阶段，我们采用确定性解码策略，以确保生成结果的稳定性与可复现性。

3)　系统性数据清洗

为缓解生成文本中固有的噪声问题，本文构建了自动评估与人工校验相结合的系统性数据清洗机制，以进一步提升数据集质量。

首先，引入基于CLIPScore (一种基于CLIP的图文一致性自动评估指标)的图文一致性自动判别模型，对每条图像－文本对进行初步筛查。的图文一致性自动判别模型，对每条图像－文本对进行初步筛查。CLIPScore通过跨模态语义相似度计算评估文本内容与图像语义的贴合程度，得分越高，表明二者的一致性越强。在完整地统计整个AGMix-PR数

据集的 CLIPScore 分布后，一个经验阈值 $\theta = 0.2$ 被引入以判定样本是否含有噪声。随后，训练集和测试集被制定了不同的清洗策略。

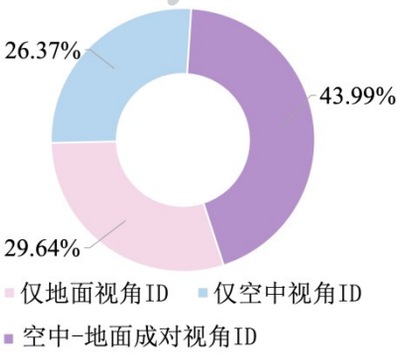


图 5　不同视角设置下的行人身份比例图

Fig. 5　The proportion of identities from different viewpoints

对于训练集来说，考虑到 AGMix-PR 数据规模庞大，且真实场景中的自然语言描述存在天然的不确定性与不精确性，训练集中适当保留噪声样本具有重要意义。一方面，可保持数据规模的完整性，避免因大规模剔除样本而导致训练数据分布受损；另一方面，适度噪声可有效提升模型在噪声环境下的稳健性，使其更能适应描述含糊、属性缺失或语义不完备等真实应用场景，降低模型对“完美标注”的依赖，减少过拟合于高度干净数据的风险。

与训练集不同，测试集构建过程中坚持严格的“零噪声”原则，以保证模型评估的客观性与可重复性。所有被自动判别为存在潜在噪声的样本均需进入人工复核环节，由具备行人检索领域经验的专业标注员完成最终确认。人工复核在参考 CLIPScore 的基础上，还着重检查文本与图像间的语义对应关系，包括但不限于：描述与视觉内容是否一致、是否存在模型幻觉、语义漂移或背景冗余、描述是否准确反映出行人的外观特征、状态与可区分细节。对于存在偏差的文本，将依据图像内容进行修订、替换或重写，确保所有测试样本均达到高度一致性与准确性。

通过上述自动与人工相结合的双阶段校验机制，最终构建的测试集在语义对齐度、内容准确性以及身份可区分性等方面均表现出高度一致性与严格性。该机制有效确保了测试数据的纯净度，为后续模型评估提供了公平、稳定且可信的实验环境，使各类方法在空－地混合视角条件下的真实性能能够得到客观、可靠的反映。图 6 给出了若干经人工修订的示例，其中针对模型生成的“wearing a red top”等与图像内容不符的细节进行了纠正；对于诸如“The image is too blurry to provide a detailed description of the person’s appearance.”等缺乏身份判别价值的描述，则通过重写增添必要的外观细节与区分信息，使文本标注更加注重细粒度的图像语义以满足行人检索任务的要求。

图 6　测试集中的人工校正的样例

Fig. 6　Manually Corrected Examples from the Test Set

经过数据清洗后，AGMix-PR 数据集共有 8 229 个身份的行人，其中 2 000 个 id 被划分为测试集，6 229 个 id 被划分为训练集。表 4 展示了行人身份在训练集与验证集中按不同视角设置的分布情况。

表 4　不同视角设置下 id 数量在训练集和测试集中的分布

Table 4　Distribution of ID number in training set and verification set under different viewpoints settings

	地面	空中	空中&地面
训练集	1 858	1 551	2 820
测试集	581	619	800

2.2　数据集的统计特性

AGMix-PR 数据集具有两个显著特性：多样化的图像视角与高质量的文本标注。

1）　多样化的图像视角

如图 5 所示，AGMix-PR 数据集中的行人 ID 根据视角覆盖情况可分为三类：空中-地面成对视角 ID，即同一行人同时包含空中与地面视角图像，占比最高，达 44.0%（共 3 620 个）；仅地面视角 ID，即该行人仅有地面摄像头拍摄的图像，约占 29.6%；以及仅空中视角 ID，即该行人仅有空中平台拍摄的图像，约占 26.4%。这种以跨视角成对样本为主体、辅以单一视角样本的数据构成方式，正是 AGMix-

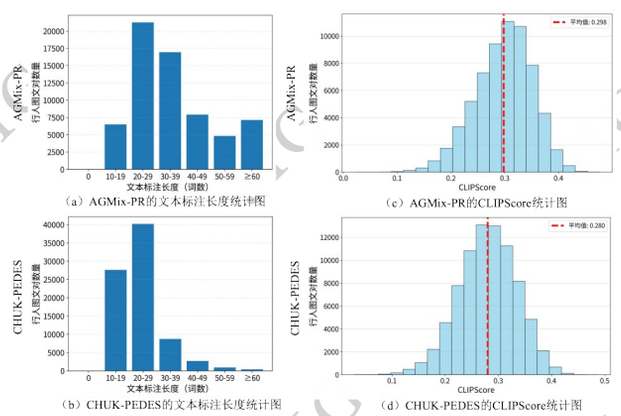
PR 的核心特征之一。引入近一半比例的单视角 ID 进一步引入了数据的多样性，真实反应了实际监控系统中视角覆盖不均衡的状况，提升模型面对不同场景的鲁棒性。

(a) AGMix-PR 词云图　　(b) CHUK-PEDES词云图

((a)word cloud of AGMix-PR dataset; (b)word cloud of CUHK-PEDES dataset)

图 7　AGMix-PR 和 CUK-PEDES 数据集词云图

Fig. 7　Word cloud of AGMix-PR and CUHK-PEDES



(a) AGMix-PR的文本标注长度统计图　(c) AGMix-PR的CLIPScore统计图

(b) CHUK-PEDES的文本标注长度统计图　(d) CHUK-PEDES的CLIPScore统计图

图 8　AGMix-PR 和 CUHK-PEDES 的文本长度分布与 CLIP-Score 分布统计图

Fig. 8　Distribution of annotation length and CLIPScore on AGMix-PR and CUHK-PEDES ((a) Statistical chart of AGMix-PR's text length distribution; (b) Statistical chart of CUHK-PEDES's text length distribution; (c) CLIPScore distribution chart of CUHK-PEDES; (d) CLIPScore distribution chart of AGMix-PR)

2)　高质量的文本标注

为全面客观地评估该数据集的文本质量，本文从词汇分布、文本粒度与 CLIPScore 分布三个方面，将其与主流数据集 CUHK-PEDES 进行对比与分析。

首先，在词汇分布方面，图 7 展示了 AGMix-PR 与 CUHK-PEDES 的词云可视化结果。可以观察到，两个数据集的高频词汇高度重合，均集中在行人描述的核心语义范畴。这表明 AGMix-PR 有效覆盖行人检索任务所需的关键属性，验证了文本的合理性。

其次，在文本长度方面，如图 8 (a)、(b)所示，CUHK-PEDES 数据集的描述长度主要集中于 10 - 29 个词，超过 60 个词的样本极为稀少；相比之下，AGMix-PR 的文本长度分布更为均衡。此外，AGMix-PR 的平均描述长度约为 39 个词，显著长于 CUHK-PEDES 的平均长度(约 23 个词)。更长的文本描述能够涵盖更多细粒度属性，从而增强语义表达的丰富性与判别能力。

最后，在图文对齐质量方面，根据图 8 (c)、(d)展示的 CLIPScore 分布，AGMix-PR 的平均分数略高于 CUHK-PEDES。这一结果表明，尽管 AGMix-PR 的文本更长、信息更密集，但并未引入语义噪声或图文不一致内容，在文本描述更加细粒度的同时保持了较高的质量。

3　评估与分析

本节采用在 TIPR 任务中常用的三种评估指标，对主流 TIPR 方法在 AGMix-PR 数据集上的性能进行系统评估，并进一步开展不同视角组合下的对比实验，以分析视角信息对模型性能的影响。

主要实验配置如下：AMD EPYC 7402 CPU(2.8 GHz，24 核)，4 NVIDIA RTX 3090 GPU(24 GB)。操作系统为 CentOS Stream 8，编程语言为 Python，深度学习框架为 Pytorch。

3.1　评估指标

本文采用行人检索的常用指标的前 k 命中率(rank-k，R-k)、平均精度均值(mean average precision，mAP)以及平均逆负样本惩罚(mean inverse negative penalty，mINP)(Ye 等，2021)来评估模型性能。

前 k 命中率用于衡量检索任务中正确匹配样本是否出现在模型返回的前 k 个结果中的比例。该指标关注模型的召回能力，即在前 k 个检索结果中是否包含与查询样本对应的真实匹配项。由于其定义简洁、直观，且贴近实际应用场景，该指标成为 TIPR 中最常用的评估指标之一。其计算公式如下：

$$\text{Rank}-k=\frac{1}{N}\sum_{i=0}^{N}\mathbb{I}(R_i\leqslant k) \tag{2}$$

式中，N 是查询样本总数，R_i 表示第 i 个查询样本对应的匹配项在检索结果列表中的排名；$\mathbb{I}(\cdot)$ 是指示函数，仅当括号内条件为真是取 1，否则为 0。k 表示在检索结果列表中的前 k 个位置，本文中分别取 1，

5，10。

平均精度均值*mAP*用于衡量模型在所有查询上的整体排序质量。与仅关注前k个结果的指标不同，*mAP*充分利用完整的检索排序列表，对每个查询计算其平均精度（average precision，AP），再对所有查询的*AP*值取平均，从而综合反映模型在精度与相关样本排名位置方面的性能。定义如下：

$$mAP = \frac{1}{N}\sum_{i}^{N} AP_i \tag{3}$$

式中，N是查询样本总数，第i个查询的AP_i为：

$$AP_i = \frac{1}{N_i}\sum_{k=1}^{n} P_i(k)\cdot rel_i(k) \tag{4}$$

式中，n代表检索返回的样本数，N_i代表第i次查询的相关样本的总数，$P_i(k)$代表前k个检索结果的精度，$rel_i(k)$为指示函数，表示第k个样本若为相关样本则为1，否则为0。

平均逆负样本惩罚是一种专为衡量模型在检索任务中定位“最难正确匹配项”所需代价而设计的评估指标。所谓“最难正确匹配项”，指的是每个查询对应的所有真实匹配样本中在排序列表里排名最靠后的一个。在空中－地面多视角混合场景下，确保目标行人在不同视角下的所有出现均不被遗漏，是衡量模型鲁棒性的关键要求；并且最难样本的排名位置直接决定了人工复核时需浏览的结果数量，进而影响实际检索效率。*mINP*通过对每个查询计算其最靠后正样本排名的倒数（Inverse Negative Penalty，INP），再对所有查询的*INP*值取平均，从而量化模型在最坏情况下的检索性能。因此，*mINP*值越高，表明模型平均只需检查更短的排序列表即可覆盖全部正确匹配，不仅显著降低人工审核成本，也反映出其在处理困难样本时具备更强的鲁棒性。具体来说，该指标定义为每个查询的*INP*的平均值为：

$$mINP = \frac{1}{N}\sum_{i=0}^{N} INP_i \tag{5}$$

式中，N是查询样本总数，第i个查询的INP_i为：

$$INP_i = \frac{|G_i|}{R_i^{hard}} \tag{6}$$

式中，R_i^{hard}代表第i个查询中最后一个正确匹配项在查询列表中的排名位置，$|G_i|$代表第i个查询的检索列表中正确匹配的总个数。

综合利用以上评测指标，本文对现有TIPR方法进行了科学、全面的测评。

3.2 基准评测

为充分评估现有TIPR方法在空中-地面多视角混合场景下的性能，本文在AGMix-PR数据集上对十种主流的TIPR方法进行了基准评测，包括：MLLM4Text（MLLMs for transferable text-to-image person reID）（Tan等，2024）、RaSa（relation and sensitivity aware representation learning method）（Bai等，2023b）、APTM（attribute prompt learning and text matching learning framework）（Yang等，2023）、AUL（adaptive uncertainty-based learning framework）（Li等，2024b）、CFine（CLIP-driven fine-grained information excavation framework）（Yan等，2023）、IRRA（cross-modal implicit relation reasoning and aligning framework）（Jiang和Ye，2023）、TBPS-CLIP（contrastive language image pretraining for TBPS）（Cao等，2024）、RDE（robust dual embedding method）（Qin等，2024）、TAG-CLIP（contrastive language image pretraining for text-based aerial-ground person retrieval）（Zhou等，2025）以及AEA-FIRM（adaptive elastic alignment network with fine-grained representation mining）（Wang等，2025a）。表5展示了上述方法在AGMix-PR与传统地面视角数据集CUHK-PEDES上的实验结果。结果表明，尽管这些方法在传统地面摄像头场景下表现良好，但在空中－地面混合视角场景中均面临显著性能下降。绝大多数方法在关键指标R-1、mAP和mINP上的性能降幅超过10%，凸显了现有TIPR方法在跨视角泛化能力方面的严重不足。

在所评测的方法中，RDE在各项评估指标上均取得最优性能，对它的进一步分析有望为多视角混合场景下的TIPR方法提供有效的优化方向。该方法针对行人重识别中普遍存在的训练数据噪声问题，即图像与文本描述之间可能存在不完美匹配或错误关联，专门设计了双重抗噪机制。具体而言，RDE通过其置信共识划分模块动态识别高可靠度的图文样本对并赋予更高权重，同时引入正则化的三元组对齐损失，有效抑制噪声样本在优化过程中的负面影响，从而显著提升模型在含噪环境下的鲁棒性。这一机制在一定程度上能够应对AGMix-PR数据集中的多视角挑战。由于同一行人身份在空中与地面视角下的图像存在显著的表征差异，而在某一特定视角下的文本描述通常在另一视角下往往难

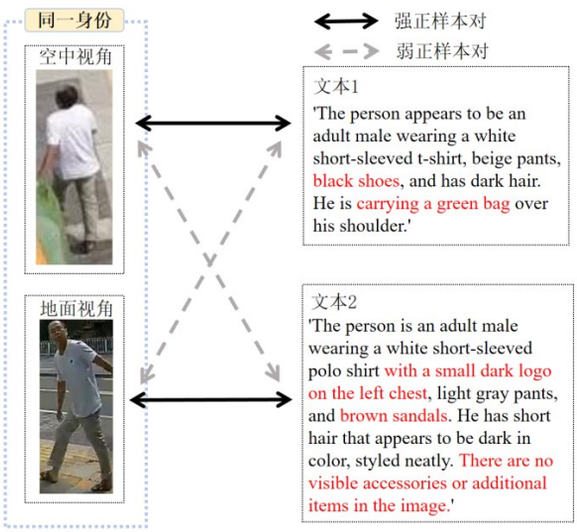


图 9　多视角行人描述一致性分析示意图

Fig. 9　Diagram of multi-view person description consistency analysis

以观测或验证，从而形成弱正样本对：属于同一行人身份但跨视角局部特征难以对齐的图文对。如图 9 所示，左侧展示了同一行人在空中与地面视角下的图像，右侧为其对应的两段文本描述（文本 1 与文本 2）。其中，基于地面视角生成的细粒度描述在空中视角下几乎不可见，导致模型难以实现精确的图文对齐。RDE 凭借其内在的抗噪能力，能够有效缓解此类由视角差异引发的学习偏差，从而实现了更鲁棒的检索性能。

3.3　文本生成方式消融实验

为分析不同文本生成方式对检索性能的影响，在 AGMix-PR 数据集上开展了文本生成策略的消融实验，实验结果如表 6 所示。当仅采用属性引导生成或仅采用模板引导生成时，模型在各项评价指标上的性能均出现不同程度的下降。该结果表明，属性引导生成与模板引导生成在语义信息覆盖与表达形式上具有较强的互补性，其联合使用能够有效提升文本描述的多样性与准确性，从而显著增强模型在检索任务中的匹配能力。

3.4　训练集保留噪声比例分析

为验证噪声数据对训练过程的影响，本节针对训练集中噪声样本的保留比例开展了系统性的消融实验。具体而言，以 CLIPScore ≤ 0.2 作为噪声样本的判定标准，在 AGMix-PR 训练集中共筛选出 2 085 个噪声样本，占训练集总量的 4.35%。在此基础上，构建了四组训练子集，分别保留原始噪声样本的 0%、33%、66% 和 100%（其余噪声样本被移除），而所有非噪声样本均完整保留。所有模型均在相同的网络架构（RDE）下进行训练，并在完全清洗的测试集上进行评估，以确保实验结果的公平性。

实验结果如表 7 所示，可以观察到当训练集中完全移除噪声样本时，模型性能相对较低；随着噪声样本逐步引入，检索性能并未出现退化；在保留全部噪声样本的设置下，模型在各项指标上均取得最优

表 5　不同 TIPR 方法在 AGMix-PR 和 CUHK-PEDES 上的检索结果

Table 5　Results of different TIPR methods on AGMix-PR and CUHK-PEDES

方法	AGMix-PR					CUHK-PEDES				
	R-1	R-5	R-10	mAP	mINP	R-1	R-5	R-10	mAP	mINP
MLLM4Text	57.27	73.84	79.86	**49.62**	–	**78.13**	**91.19**	94.50	68.75	–
RaSa	61.46	77.71	83.45	48.94	**30.27**	76.51	90.29	94.25	**69.38**	–
APTM	61.62	77.73	83.54	47.72	27.70	76.53	90.04	94.15	66.91	–
CFine	56.82	74.83	81.46	42.09	15.38	69.57	85.93	91.15	–	–
IRRA	60.10	77.73	83.69	47.23	27.51	73.38	89.93	93.91	66.13	50.24
TBPS-CLIP	60.22	77.31	83.01	46.49	26.52	73.54	88.19	92.35	65.38	–
RDE	**61.88**	**78.29**	**83.98**	47.97	27.76	75.94	90.14	94.12	67.56	**51.44**
AUL	59.71	76.42	82.66	46.03	26.03	69.16	83.32	88.37	–	–
TAG-CLIP	60.66	77.70	83.31	48.04	–	74.38	88.30	92.59	67.18	–
AEA-FIRM	52.28	71.03	80.11	40.55	–	76.54	90.79	**94.51**	68.10	–

注：黑色加粗字体表示最优结果，下划线表示次优结果。

表 6　文本生成方式消融实验

Table 6　Ablation of text generation

方法	R-1	R-5	R-10	mAP	mINP
仅属性引导生成	41.78	58.16	64.55	31.20	15.49
仅模板引导生成	44.07	58.94	66.46	33.29	16.83
属性引导生成+模板引导生成	**60.22**	**77.31**	**83.01**	**46.49**	**26.52**

注：黑色加粗字体表示最优结果。

表现，显示出适度噪声对模型学习的潜在正向作用。这一现象表明，当前训练集中噪声样本并未破坏整体语义分布，反而可能作为一种隐式正则化，引入轻微扰动，从而缓解过拟合并提升模型的泛化能力。基于上述分析，最终实验中选择保留训练集中真实存在的低置信度样本，而非对训练数据进行过度清洗。该策略既有助于保持数据分布的真实性，也更符合复杂空－地多视角场景下实际数据噪声客观存在的特点。

3.5　不同视角组合的对比分析

为了分析不同视角组合数据对于模型检索性能影响，本节在AGMix-PR数据集中选取了3 620个同时拥有地面和空中视角图像的行人身份，其中2 820个身份作为训练集，800个身份作为测试集。在此基础之上，构建了三种不同的训练数据组合：1）仅地面视角：使用训练集中所有身份的地面视角图像；2）仅空中视角：使用训练集中所有身份的空中视角图像；3）空中-地面混合视角：为每个行人随机选取3张图像（可能包含地面、空中或两者混合），确保三种组合的训练集身份数量完全一致。

表 7　保留不同噪声比例对模型训练效果的对比实验

Table 7　Comparative experiment of on the effect of retaining different noise ratios

噪声比例	方法	R-1	R-5	R-10	mAP	mINP
0%	RDE	60.82	76.90	82.77	47.05	27.32
33%	RDE	60.66	77.01	82.74	47.09	27.16
66%	RDE	61.71	77.96	82.76	47.24	27.42
100%	RDE	**61.88**	**78.29**	**83.98**	**47.97**	**27.76**

注：黑色加粗字体表示最优结果。

如表8所示，本节选取了在基准测评中的4种方法（APTM、IRRA、TBPS-CLIP和RDE）进行不同视角组合对比实验。4种方法在空中-地面混合组合下均表现出最佳的检索性能，由此可见，空中-地面混合数据训练能普遍提升模型的鲁棒性和泛化能力。与此同时，RDE方法在三种数据组合下均稳居第一，尤其是在挑战性更高的仅空中视角设置下依然保持显著优势，再次印证了其噪声鲁棒性设计在处理视角相关噪声方面的有效性。除此之外，APTM方法依托大规模合成的MALS数据集，将行人属性提示与文本匹配任务进行联合建模，借助属性先验显式强化细粒度图文对齐，从而为模型提供更强的先验表征能力，使其在空-地混合数据训练场景下同样保持领先的检索性能。

表 8　不同视角组合分析实验结果

Table 8　Analysis of experimental results from different perspective combinations

方法	仅地面视角					仅空中视角					空中-地面混合视角				
	R-1	R-5	R-10	mAP	mINP	R-1	R-5	R-10	mAP	mINP	R-1	R-5	R-10	mAP	mINP
APTM	52.08	70.10	77.15	35.09	12.59	50.37	69.87	77.35	33.96	10.59	**57.72**	**75.15**	**81.79**	**40.24**	**16.33**
IRRA	50.69	67.79	77.63	34.79	13.06	47.65	68.41	76.98	33.51	11.92	**55.83**	**75.10**	**81.66**	**39.32**	**16.23**
TBPS-CLIP	50.84	69.87	77.17	33.28	13.21	48.36	69.17	77.08	33.28	11.24	**55.48**	**74.24**	**81.06**	**38.45**	**15.60**
RDE	52.86	70.82	77.95	36.31	13.95	51.89	71.05	78.78	36.17	13.19	**58.70**	**76.75**	**82.97**	**41.06**	**16.87**

注：黑色加粗字体表示最优结果。

4　结　论

本文提出了一种文本-图像的空中-地面多视角混合行人检索数据集，旨在为低空经济大背景下兴起的空中-地面多视角混合场景提供全新的基准数据集。一方面，AGMix-PR系统地整合了无人机视角与地面摄像头视角下的大规模行人图像及其文本描述，为研究跨视角、跨模态的统一表征学习提供了基础数据支持。方法上，本文围绕行人场景的细粒度

描述需求,构建了属性引导的层次化文本生成框架,将多模态大模型的语义生成能力与行人领域的结构化先验(如服饰类别、颜色、携带物品等)相结合,在显著降低人工标注成本的同时,保证生成文本在可读性与辨识度上的均衡。实验上,在 AGMix-PR 上全面评估了多种代表性文本-图像的行人检索方法,覆盖从传统双编码器架构到视觉-语言预训练模型在内的多种范式,系统分析了不同方法在空视角、地视角以及空-地混合检索条件下的性能表现与鲁棒性差异。基于这些结果,本文进一步总结了当前技术在空-地多视角场景下面临的核心瓶颈,并给出相应的研究启示,为后续方法设计与系统落地提供经验参考与改进方向。

在此基础上,本文结合 AGMix-PR 的构建过程与实验分析,对文本-图像的空中-地面多视角行人检索任务所面临的新挑战进行了梳理与讨论,主要体现在以下四个方面:

1)视角信息建模与解耦。空中与地面视角在成像高度、俯仰角、人物尺度与背景结构等方面存在显著差异,同一行人跨视角外观变化剧烈,导致特征分布偏移。如何在统一嵌入空间中合理编码并部分解耦视角因素,使模型既能利用视角信息又能稳定捕捉与身份相关的服饰与属性语义,是空-地多视角混合行人检索面临的核心挑战之一。

2)弱正样本学习与样本采集机制。在多视角场景下,文本在身份层面虽与行人匹配,但细粒度语义往往仅在特定视角可见,形成难以与所有图像严格对齐的弱正样本,例如地面视角的鞋款细节在无人机视角中不可辨识。如何在弱正样本和标注噪声普遍存在的条件下,稳健学习身份一致表征,并与真实采集与标注流程相协调,是空中-地面多视角混合行人检索需要重点关注的问题。

3)运动感知与时序信息建模。空-地协同监控中,无人机巡航与摄像头跟踪使行人观测过程天然具有时序性与运动性。如何将动作模式、姿态变化与轨迹等动态线索与静态外观特征有效融合,构建面向“运动中的行人”的文本检索模型,仍有较大研究空间。

4)轻量化模型设计与部署友好性。无人机与边缘节点受算力、存储与能耗限制,难以直接部署大规模视觉-语言模型。如何在有限资源下兼顾检索精度与推理效率,设计结构紧凑、计算友好的轻量化行人检索模型,是推动实际落地亟需解决的问题。

参考文献(References)

Bai J Z, Bai S, Yang S S, Wang S J, Tan S N, Wang P, et al. 2023a. Qwen-vl: a frontier large vision-language model with versatile abilities[EB/OL].[2023-10-13]. https://arxiv.org/pdf/2308.12966.pdf.

Bai Y, Cao M, Gao D M, Cao Z Q, Chen C, Fan Z F, et al. 2023b. Rasa: relation and sensitivity aware representation learning for text-based person search[EB/OL]. [2023-05-23]. https://arxiv.org/pdf/2305. 13653.pdf.

Cao M, Bai Y, Zeng Z Y, Ye M and Zhang M.2024. An empirical study of clip for text-based person search//Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver, British Columbia, Canada: AAAI press: 465-473 [DOI: 10.1609/aaai. v38i1. 27801]

Cao M, Zhou X Y, Jiang D, Du B, Ye M and Zhang M. 2025.Multilingual text-to-image person retrieval via bidirectional relation reasoning and aligning. IEEE transactions on pattern analysis and machine intelligence, N/A(N/A): 1-18 [DOI: 10.1109/TPAMI. 2025. 3620139]

Chen Y C, Huang R, Chang H, Tan C Q, Xue T and Ma B P. 2021. Cross-modal knowledge adaptation for language-based person search. IEEE Transactions on Image Processing, 30(N/A): 4057-4069 [DOI: 10.1109/TIP.2021.3068825]

Ci Y Z, Wang Y Z, Chen M L, Tang S X, Bai L, Zhu F, et al.2023. Unihcp: a unified model for human-centric perceptions//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Vancouver, BC, Canada: IEEE: 17840-17852 [DOI: 10. 1109/CVPR52729.2023.01711]

Ding Z, Ding C, Shao Z and Tao D. 2021. Semantically self-aligned network for text-to-image part-aware person re-identification [EB/OL].[2021-07-27]. https://arxiv.org/pdf/2107.12666.pdf.

Dosovitskiy A. 2020. An image is worth 16x16 words: Transformers for image recognition at scale[EB/OL]. [2020-10-22]. https://arxiv.org /pdf/2010.11929.pdf.

Graves A. 2012. Long short-term memory. Supervised sequence labelling with recurrent neural networks, 385(N/A): 37-45[DOI: 10. 1007 /978-3-642-24797-2]

Jiang D and Ye M. 2023. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Vancouver, BC, Canada: IEEE: 2787-2797 [DOI: 10.1109/CVPR 52729.2023.00273]

Jiang J Y, Ding C X, Tan W T, Wang J H, Tao J and Xu X M.2025. Modeling thousands of human annotators for generalizable text-to-

image person re-identification//Proceedings of the Computer Vision and Pattern Recognition Conference. Nashville, TN, USA: IEEE: 9220-9230 [DOI: 10.1109/CVPR52734.2025.00861]

Jing Y, Si C Y, Wang J B, Wang W, Wang L and Tan T N. 2020. Pose- guided multi-granularity attention network for text-based person search//Proceedings of the AAAI conference on artificial intelli- gence. Palo Alto, California USA: AAAI press: 11189-11196 [DOI: 10.1609/aaai.v34i07.6777]

Leng J X, Mo M J C, Zhou Y H, Ye Y M, Gao C Q and Gao X B. 2023. Recent advances in drone-view object detection. Chinese Journal of Image Graphics, 28(9): 2563-2586 (冷佳旭, 莫梦竞成, 周应华, 叶永明, 高陈强, 高新波. 2023. 无人机视角下的目标检测研究进展. 中国图象图形学报, 28(9): 2563-2586) [DOI: 10.11834/jig.220836].

Li B, Zhang Y H, Guo D, Zhang R R, Li F, Zhang H, et al. 2024. Llava- onevision: easy visual task transfer [EB/OL]. [2024-08-06].
https:// arxiv.org/pdf/2408.03326.pdf.

Li J N, Selvaraju R, Gotmare A, Joty S, Xiong C M and Hoi S C H. 2021a. Align before fuse: vision and language representation learning with momentum distillation. Advances in neural information processing systems, 34 (N/A): 9694-9705 [DOI: 10.48550/arXiv. 2107.07651]

Li N N and Chen Z Z.2018.Image cationing with visual-semantic LSTM // Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence. Stockholm, Sweden: International Joint Conferences on Artificial Intelligence Organization: 793-799 [DOI: 10.24963/ijcai.2018/110]

Li S, Xiao T, Li H, Zhou B, Yue D and Wang X.2017a. Person search with natural language description//Proceedings of the IEEE conference on computer vision and pattern recognition. Honolulu, HI, USA:IEEE:1970-1979 [DOI: 10.1109/CVPR.2017.551]

Li S, Xiao T, Li H S, Yang W and Wang X G. 2017b. Identity-aware text-ual-visual matching with latent co-attention//Proceedings of the IEEE international conference on computer vision. Venice, Italy: IEEE:1890-1899 [DOI: 10.1109/ICCV.2017.209]

Li T J, Liu J, Zhang W, Ni Y and Li Z H. 2021b. UAV-Human: a large benchmark for human behavior understanding with unmanned aerial vehicles//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, TN, USA: IEEE [DOI: 10.1109/ CVPR46437.2021.01600]

Nguyen H, Nguyen K, Sridharan S and Fookes C. 2023. Aerial-ground person re-id//2023 IEEE International Conference on Multimedia and Expo (ICME). Brisbane, Australia: IEEE: 2585-2590 [DOI: 10. 1109/ICME55011.2023.00440]

Nguyen H, Nguyen K, Sridharan S and Fookes C. 2024. AG-ReID. v2: bridging aerial and ground views for person re-identification. IEEE Transactions on Information Forensics and Security, 19 (N/A): 2896 -2908 [DOI: 10.1109/TIFS.2024.3353078]

Qin Y, Chen Y K, Peng D Z, Peng X, Zhou J T and Hu P. 2024. Noisy- correspondence learning for text-to-image person re-identification //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE: 27197-27206 [DOI: 10.1109/CVPR52733.2024.02568]

Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. 2021. Learning transferable visual models from natural language supervision//International conference on machine learning. virtual conference: PmLR: 8748-8763 [DOI: 10.48550/arXiv. 2103.00020]

Simonyan K and Zisserman A. 2014. Very deep convolutional networks for large-scale image recognition [EB/OL]. [2014-09-04]. https:// arxiv.org/pdf/1409.1556.pdf.

Song Z F, Hu G S and Zhao C R.2024. Diverse person: Customize your own dataset for text-based person search//Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver, British Columbia, Canada: AAAI Press: 4943-4951 [DOI: 10.1609/aaai. v38i5.28298.]

Tan W T, Ding C X, Jiang J Y, Wang F, Zhan Y B and Tao D P.2024. Harnessing the power of mllms for transferable text-to-image person reid//Proceedings of the IEEE/CVF Conference on Com- puter Vision and Pattern Recognition. Seattle, WA, USA: IEEE: 17127-17137 [DOI: 10.1109/CVPR52733.2024.01621]

Team Q. 2024. Qwen2 technical report[EB/OL].[2024-07-15]. https:// arxiv.org/pdf/2407.10671.pdf.

Vinyals O, Toshev A, Bengio S and Erhan D. 2016. Show and tell: lessons learned from the 2015 mscoco image captioning challenge. IEEE transactions on pattern analysis and machine intelligence, 39 (4): 652-663 [DOI: 10.1109/TPAMI.2016.2587640]

Wang W Y, Chen Z, Wang W H, Cao Y, Liu Y Z, Gao Z W, et al. 2024. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization [EB/OL]. [2024-11-15].
https://arxiv.org/pdf/2411.10442.pdf.

Wang X C, Han Y Q, TANG L B and DENG C W. 2022.Multi target detection and tracking algorithm for UAV platform based on deep learning. Journal of Signal Processing, 38(1): 157-163 (王旭辰, 韩煜祺, 唐林波, 邓宸伟. 2022. 基于深度学习的无人机载平台多目标检测和跟踪算法研究. 信号处理, 38(1): 157-163) [DOI: 10. 16798/j.issn.1003-0530.2022.01.018]

Wang Y H, Yang M, Cao R and Gao G W. 2025b. AEA-FIRM: adaptive elastic alignment with fine-grained representation mining for text- based aerial pedestrian retrieval. IEEE Transactions on Circuits and Systems for Video Technology, N/A(N/A): 1-1 [DOI: 10.1109/ TCSVT.2025.3586601]

Xiao Y, Cao D, Li C L, Jiang B and Tang J. 2025.A benchmark dataset for high-altitude UAV multi-modal tracking. Chinese Journal of Image Graphics, 30(2): 361-374 (肖云, 曹丹, 李成龙, 江波, 汤进. 2025. 基于高空无人机平台的多模态跟踪数据集. 中国图

象图形学报，30(2)：361-374)[DOI:10.11834/jig.240040]

Xu Y J，Wu M X，Guo Z X，Cao M，Ye M and Laaksonen J. 2025. Efficient text-to-video retrieval via multi-modal multi-tagger derived pre-screening. Visual Intelligence，3(1)：1-13 [DOI：10.1007/s44267-025-00073-2]

Yan S L，Liu J，Dong N，Zhang L Y and Tang J H.2024. Prototypical prompting for text-to-image person re-identification//Proceedings of the 32nd ACM International Conference on Multimedia. New York，NY，USA:ACM:2331-2340 [DOI：10.1145/3664647.3681165]

Yang S Y，Zhou Y N，Zheng Z D，Wang Y X，Zhu L and Wu Y J.2023. Towards unified text-based person retrieval：a large-scale multi-attribute and language search benchmark//Proceedings of the 31st ACM international conference on multimedia. New York，NY，USA:ACM:4492-4501 [DOI：10.1145/3581783.3611709]

Ye M，Shen J B，Lin G J，Xiang T，Shao L and Hoi S C. 2021. Deep learning for person re-identification：A survey and outlook. IEEE transactions on pattern analysis and machine intelligence，44(6)：2872-2893 [DOI：10.1109/TPAMI.2021.3054775]

Zhai X H，Mustafa B，Kolesnikov A and Beyer L.2023. Sigmoid loss for language image pre-training//Proceedings of the IEEE/CVF international conference on computer vision. Paris，France：IEEE：11975- 11986 [DOI：10.1109/ICCV51070.2023.01100]

Zhang Q，Wang L，Patel V M，Xie X H and Lai J H.2024. View-decoupled transformer for person re-identification under aerial-ground camera network//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle，WA，USA：IEEE:22000-22009 [DOI：10.1109/CVPR52733.2024.02077]

Zhang S Z，Yang Q C，Cheng D，Xing Y H，Liang G Q，Wang P and Zhang Y N. 2023. Ground-to-aerial person search：benchmark dataset and approach//Proceedings of the 31st ACM International Conference on Multimedia. New York，NY，USA:ACM:789-799 [DOI：10.1145/3581783.3612105]

Zhang S Z，Zhang Q，Yang Y F，Wei X，Wang P，Jiao B L and Zhang Y N. 2020. Person re-identification in aerial imagery. IEEE Transactions on Multimedia，23(N/A)：281-291 [DOI：10.1109/TMM.2020.2977528]

Zhang X Y，Sun X S，Luo Y P，Ji J Y，Zhou Y Y，Wu Y J，et al.2021. RSTNet：captioning with adaptive attention on visual and non-visual words//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Nashville，TN，USA:IEEE：15465-15474 [DOI：10.1109/CVPR46437.2021.01521]

Zheng L，Shen L Y，Tian L，Wang S J，Wang J D and Tian Q. 2015. Scalable person re-identification：a benchmark//Proceedings of the IEEE international conference on computer vision. Santiago，Chile：IEEE:1116-1124 [DOI：10.1109/ICCV.2015.133]

Zhou X Y，Wu Y，Ma J Y，Wang W H，Cao M and Ye M. 2025. Text-based aerial-ground person retrieval[EB/OL].[2025-11-11]. https: //arxiv.org/pdf/2511.08369.pdf.

Zhu A C，Wang Z J，Li Y F，Wan X L，Jin J，Wang T，et al. 2021. DSSL：deep surroundings-person separation learning for text-based person retrieval [EB/OL]. [2021-09-12]. https://arxiv.org/pdf/2109.05534. pdf.

Zhu D Y，Chen J，Haydarov K，Shen X Q，Zhang W X and Elhoseiny M. 2023. Chatgpt asks，blip-2 answers：automatic questioning towards enriched visual descriptions[EB/OL].[2023-03-12]. https://arxiv.org /pdf/2303.06594.pdf.

Zuo J L，Zhou H Y，Nie Y，Zhang F，Guo T Y，Sang N，et al.2024. Ufinebench：towards text-based person retrieval with ultra-fine granularity//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle，WA，USA：IEEE：22010-22019 [DOI：10.1109/CVPR52733.2024.02078]

作者简介

周新宇，男，博士研究生，主要研究方向为行人重识别和跨模态检索。E-mail：xyzhou25@stu.suda.edu.cn

曹敏，通信作者，女，副教授，主要研究方向为行人重识别和跨模态视觉-语言学习。E-mail：caomin0719@126.com

吴彧，男，硕士研究生，主要研究方向为行人重识别和多模态推理。E-mail：20235227114@stu.suda.edu.cn

朱巧明，男，教授，主要研究方向为自然语言处理和知识图谱。E-mail：qmzhu@suda.edu.cn

叶茫，男，教授，主要研究方向为多模态检索与生成和联邦学习。E-mail：yemang@whu.edu.cn