AG-VPReID.VIR: 为
基于视频的可见光-红外人员重识别搭建空中与地面平台的桥梁

Huy Nguyen, Kien Nguyen, Akila Pemasiri, Akmal Jahan, Clinton Fookes, and Sridha Sridharan
School of Electrical Engineering and Robotics, Queensland University of Technology
{ nguyet91, k.nguyenthanh, a.thondilege, akmal.jahan, c.fookes, s.sridharan } @qut.edu.au

Abstract

跨可见光和红外模态的人重新识别（Re-ID）对于 24 小时监控系统至关重要，但现有数据集主要集中在地面视角。虽然地面红外系统提供夜间能力，但它们受到遮挡、覆盖范围有限和对障碍物的脆弱性等问题的困扰——而这些问题可以通过空中视角独特地解决。为了解决这些限制，我们引入了 AG-VPReID.VIR，这是首个空地跨模态基于视频的人重新识别数据集。该数据集使用安装在无人机和固定的闭路电视摄像头的 RGB 和红外模态，捕捉了 1,837 个身份，跨越 4,861 个轨迹（124,855 帧）。AG-VPReID.VIR 呈现了独特的挑战，包括跨视角变化、模态差异和时间动态。此外，我们提出了 TCC-VPReID，一种新颖的三流架构，旨在应对跨平台和跨模态人重新识别的共同挑战。我们的方法通过风格鲁棒的特征学习、基于记忆的跨视角适应和中介引导的时间建模，弥合了空地视角和 RGB-IR 模态之间的域差距。实验表明，AG-VPReID.VIR 与现有数据集相比提出了独特的挑战，我们的 TCC-VPReID 框架在多个评估协议中实现了显著的性能提升。数据集和代码可以在 https://github.com/agvpreid25/AG-VPReID.VIR 获得。

1. 引言

行人再识别（Re-ID）旨在匹配在不同相机、时间或视角下捕获的个人图像或视频。在监控和安全系统中，它已成为计算机视觉中的一项关键任务 [26, 38] 。传统的再识别方法主要集中在匹配从地基相机在光线良好的条件下获取的 RGB 图像 [18, 14] 。然而，现实世界的监控场景通常涉及不同的照明条件，包括低光或夜间环境，其中 RGB 相机可能无法捕捉到有用的信息 [28, 16] 。在这种情况下，红外（IR）成像由于其在低光条件下捕捉图像的能力提供了一种可行的解决方案 [4, 15] 。

在行人重识别领域，一些研究已经探索了空地设置。Nguyen 等人引入了具有 1,615 个身份的 100,502 张 RGB 图像的 AG-ReID.v2，而 Zhang 等人提出了来自 2,788 个身份的 185,907 帧视频的 G2A-VReID。然而，这两者都仅限于 RGB 模式。尽管在地面摄像机的跨模态重识别中取得了显著进展，结合空中与地面视角的 RGB-IR 模式仍未被探索。跨模态行人重识别，特别是在 RGB 和 IR 图像之间，引起了极大兴趣，因为它通过弥合昼夜差距实现了 24 小时监控。已经提出了几个用于跨模态重识别的数据集，如 SYSU-MM01、RegDB、HITSZ-VCM 和 BUPTCampus。然而，这些数据集主要关注基于地面的监控，并未考虑空地设置的跨模态场景，这是涉及无人机和地面摄像机的综合监控系统所必需的。

为了解决这些限制，我们引入了 AG-VPReID.VIR，一个新颖的用于空地跨模态视频的行人重识别数据集。该数据集通过结合从空中和地面视角捕获的 RGB 和 IR 模态扩展了现有数据集。跨平台（空中与地面）和跨模态（RGB vs. IR）的综合挑战提出了需要解决的独特问题。该数据集包含来自 1,837 个身份的 4,861 条轨迹，由多台无人机和固定闭路电视摄像机在两种模态下捕获。这一多样化的集合呈现了包括视点变化、模态差异以及时间动态的独特挑战，使其成为开发鲁棒重识别系统的宝贵基准。图 ?? 展示了我们提出的数据集及其与之前提出的设置的区别。

此外，我们提出了 TCC-VPReID（三流架构用于跨平台交叉模态视频的行人重识别），这是一个专为 RGB-IR 空地行人重识别设计的新颖架构。我们的方法旨在有效利用空间和时间信息，同时解决空地视角和 RGB-IR 模态之间的显著域差距。通过对 AG-VPReID.VIR 进行广泛实验，我们证明了我们的方法相比现有的最先进方法取得了更优越的性能。

本研究的主要贡献包括：(1) 我们介绍了首个基于视频的空地交叉模态行人重识别数据集，从而推动了对更实际的监控系统的研究；(2) 我们通过对最先进方法的全面基准测试建立了基线性能指标；(3) 我们开发了 TCC-VPReID，这是一个三流架构，专门设计用于通过风格鲁棒特征学习、基于记忆的跨视角适应和时序建模来解决空地视角变化和 RGB-IR 模态差异。数据集、基线模型以及我们的方法都将公开发布。

可见-红外行人重新识别数据集：RGB-红外行人重新识别因全天候监控需求而受到关注。SYSU-MM01 [28] 首创了包含 491 个身份的跨 6 个摄像头的数据集，而 RegDB [21] 提供了配对的 RGB-热成像的 412 个身

Table 1. AG-VPReID.VIR 与其他行人再识别数据集的比较。

Dataset	View & Modality				Statistics		
	Ground		Aerial				
	RGB	IR	RGB	IR	# IDs	# Tracklets	# Frames
Image-based Datasets							
AG-ReID.v2 [22]	✓	-	✓	-	1,615	-	100,502
RegDB [21]	✓	✓	-	-	412	-	8,240
SYSU-MM01 [28]	✓	✓	-	-	491	-	303,420
LLCM [42]	✓	✓	-	-	1,064	-	46,767
Video-based Datasets							
G2A-VReID [41]	✓	-	✓	-	2,788	5,576	180,000
HITZS-VCM [16]	✓	✓	-	-	927	21,863	463,259
BUPTCampus [6]	✓	✓	-	-	3,080	16,826	1,869,366
AG-VPReID.VIR (Ours)	✓	✓	✓	✓	1,837	4,861	124,855

份。视频数据集则包括 HITSZ-VCM [16] （927 个身份，463,000 帧）和 BUPTCampus [6] （3,080 个身份，187 万帧）。然而，这些数据集仅关注地面视角，缺乏对于全面监控至关重要的空中视角。表 1 和图 1 将我们提出的数据集与其他可见-红外数据集进行了比较。

航空-地面行人重识别数据集：重识别研究中的航空-地面视角随着 AG-ReID.v2 [22] （100,502 张 RGB 图像，1,615 个身份）和 G2A-VReID [41] （185,907 个视频帧，2,788 个身份）而出现。然而，这些数据集仅包含可见光谱影像，限制了夜间的实用性。我们的数据集独特地结合了两种模式下的航空和地面视图，以实现跨多种光照和视角的稳健重识别。

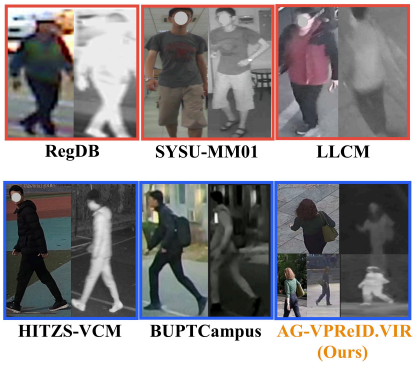

Figure 1. 我们可见光-红外图像 vs. 现有数据集的可见光-红外图像。红色边框 表示基于图像的对比，蓝色边框 表示基于视频的设置。

可见光-红外行人重识别：跨模态行人重识别方法主要分为两大类：通过专门的架构学习模态共享特征 [35, 34] ，以及利用对抗学习 [4, 27] 来对齐模态之间的特征分布。大多数现有的方法假设训练和测试阶段的数据分布相同——但在实际的重识别场景中，身份集通常不重叠，这一假设不成立。此外，这些方法主要考虑来自相似地面视角的图像，忽略了显著的视点变化。我们的工作特别解决了跨模态和跨平台变化的联合挑战，这代表了现有文献中的一个重要空白。

基于视频的人重识别：基于视频的人重识别方法通过三种主要方法提取时间信息：使用循环神经网络 [20, 32, 19] 进行序列依赖建模，使用时间池化操作 [32, 3, 31] 进行帧级特征聚合，以及使用 3D 卷积 [13, 10, 19] 进行直接的时空模式提取。最近的进展包括时空记忆网络 [7, 16, 6] ，它存储空间干扰和时间模式的表示。这些方法虽然在单一领域设置中表现成功，但通常分别解决跨模态挑战 [16, 6] 或跨视角变化 [41] 。我们的工作旨在通过一种新颖的架构，克服这些前所未有的领域间隔，这种架构针对从空中到地面的 RGB-IR 视频重识别进行了优化。

2. AG-VPReID.VIR 数据集

2.1. 数据集收集

AG-VPReID.VIR 数据集是在一个大学校园中通过结合无人机、固定的闭路电视摄像机和可穿戴设备收集的。受到先前发布的数据集 [22] 的启发，该数据集包含了来自无人机、闭路电视和可穿戴摄像头的 RGB 图像，AG-VPReID.VIR 通过整合无人机和配备热传感器的闭路电视摄像机捕获的红外图像来增强数据多样性。此补充使得在各种光照条件下实现全面的监控能力成为可能，特别是在 RGB 摄像机效果不佳的低光或夜间环境中。

对于无人机平台，我们使用了搭载双传感器相机的 DJI M600 Pro 无人机，能够同步捕获 RGB 和红外（IR）图像。无人机在 15 到 45 米的高度范围内操作，从高处以不同的分辨率和比例拍摄个体。对于地面监控，我们使用了两台闭路电视摄像机，包括一台标准的可见光（RGB）摄像机和一台热成像（IR）摄像机。可穿戴设备提供第一人称视角的地面 RGB 图像。表格 ?? 提供了我们摄像机设置的详细规格。摄像机位置和视野覆盖范围在附录章节 5 中展示。

数据收集在五个月内进行，涵盖了 10 个不同的时段。在每个时段内，无人机在不同高度（15 米、25 米和 45 米）进行飞行，每次飞行持续 15 分钟，以确保全面覆盖。我们总共收集了 124,855 张涉及 1,837 个独特身份的图像。AG-VPReID.VIR 数据集包括在不同相机模式和视角下捕获的 RGB 和红外图像。红外图像的包含显著增强了数据集在不同环境条件下进行行人再识别任务的适用性。

为了确保准确的跨模态和跨平台人员重新识别，数据集中的每个个体在 RGB 和 IR 图像上以及在无人机、闭路电视和可穿戴摄像头视角下都被人工标注了一个唯一的身份标签。最初，使用了 YOLOv10x(Ultralytics) 检测器及其预训练权重作为骨干进行人员检测和跟踪，并定期提取帧。标注人员手动审核自动检测到的轨迹，以纠正任何不准确之处，确保每个轨迹对应于所有模态和平台上单一的个体。这个两阶段的过程在保持标注工作流效率的同时，确保了高质量的标注。

最终的数据集包含 124,855 张图像，涵盖 5 个相机视角下捕获的 1,837 个独特身份。标注过程考虑到了诸如图像分辨率变化、显著的视角差异以及 RGB 和 IR 图像之间的模式差异等挑战。多个标注者交叉验证了标签以提高标注准确性。数据集分布分析在附录第 6 节中呈现。

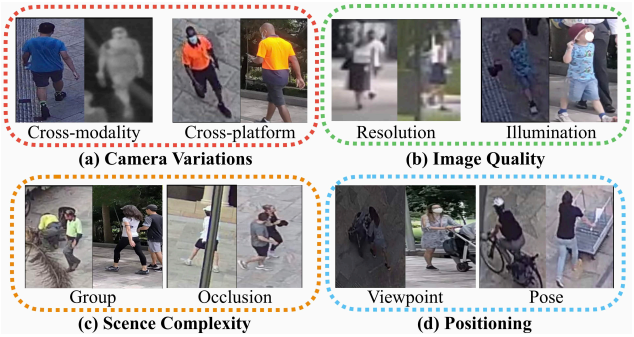

Figure 2. 我们提出的数据集中的挑战。

2.2. 数据集特征

AG-VPReID.VIR 数据集相比现有的 RGB-IR 人重新识别数据集，如 RegDB [21] 、SYSU-MM01 [28] 、HITSZ-VCM [16] 和 BUPTCampus [6] ，展现了几个独特的特性。表 1 提供了这些数据集的对比概述。AG-VPReID.VIR 数据集的关键特性包括：

- 独特的多平台多模态集成：与现有的数据集不同，这些数据集要么仅捕获地面 RGB-IR 数据（如 SYSU-MM01, RegDB, HITSZ-VCM, BUPTCampus），要么仅捕获跨空中-地面平台的 RGB 数据（如 AG-ReID.v2, G2A-VReID），我们的数据集独特地结合了来自空中和地面相机的 RGB 和 IR 模态，首次为行人重识别提供了全面的跨平台跨模态集合。

- 独特的空中 IR：我们的数据集是第一个引入来自无人机平台的 IR 图像用于行人重识别的。如图 2 (a) 所示，与地面 IR 图像相比，空中 IR 图像非常具有挑战性，为研究界带来了有趣的挑战。

- 新视角和尺度挑战：该数据集从不同角度和距离捕捉个体，导致外观和尺度的显著变化。由于不同的相机高度和类型，UAV 捕获的图像大小从 31×59 到 371×678 像素，而地面基础上的图像则从 22×23 到 172×413 像素，如图 2 (a) 和图 2 (b) 所示。

- 其他挑战：此外，我们的数据集在人员重识别中提出了各种挑战，包括如图 2 (c) 所示的场景复杂性，如部分遮挡的目标和成群行走的个体，以及如图 2 (d) 所示的不同视角和姿态带来的定位挑战。

2.3. 伦理与隐私

本研究已获得机构审查委员会的全面伦理批准。所有参与者在数据收集前均提供了知情同意书。为了保护隐私，我们采用了“去面件”[5] 进行面部匿名化，实施了访问受限的安全数据存储协议，并建立了负责任的数据使用明确指引。

3. 我们的方法

我们提出了 TCC-VPReID，这是一个新颖的三流框架，专门设计用于解决跨平台跨模态人员重识别的独特挑战。具体而言，我们的方法设计了互补技术以处理 (1) 平台间的风格变化和外观失真（流 1），(2) 基于视频的时间和跨视角的挑战（流 2），以及 (3) 可见光域与红外域之间的模态差异挑战（流 3）。图 3 展示了我们的架构，突出显示了三个专业流和特征融合模块，该模块可保留互补信息，同时协调平台和模态之间的矛盾。

3.1. 流 1：风格鲁棒特征学习

为了克服空中地面平台和 RGB-IR 模式之间的显著风格变化和外观失真，我们的第一条流实现了一种新颖的风格鲁棒特征提取机制。虽然我们受到最近风格增强技术 [45] 的启发，但我们显著扩展了这种方法，通过平台特定的风格转换，针对空中和地面视角的独特特征进行定制。

此流通过策略性地改变输入帧风格生成多样化的训练表示：

$$I_t^{aug} = [\alpha I_t^R, \beta I_t^G, \gamma I_t^B], \tag{1}$$

$$I_t^{ir,aug} = \delta I_t^{ir}, \tag{2}$$

其中 I_t 和 I_t^{ir} 分别表示第 t 个 RGB 和红外帧。I_t^R 、I_t^G 、I_t^B 表示 I_t 的红、绿和蓝通道，α 、β 、γ 和 δ 是从均匀分布 $U(0.5, 1.5)$ 中采样的，如 [45] 中所提出的。我们的创新在于开发了一种跨平台的防御机制，该机制在极端视角和模态变化情况下保持特征一致性。我们在网络的卷积块之间战略性地放置了模态内风格攻击：

$$\tilde{f}_{conv3}^{i,j} = \frac{\sigma(f_{conv3}^{k,l})}{\sigma(f_{conv3}^{i,j})} \cdot (f_{conv3}^{i,j} - \mu(f_{conv3}^{i,j})) + \mu(f_{conv3}^{k,l}), \tag{3}$$

其中 $f_{conv3}^{i,j}$ 表示 i 个身份和 j 个帧的第三卷积块特征，$f_{conv3}^{k,l}$ 表示选定用于干扰的不同身份-帧对（k , l ）的特征，$\sigma(\cdot)$ 和 $\mu(\cdot)$ 表示输入特征的方差和均值操作。此方法迫使模型通过我们的专门优化目标开发视角不变和模态不变的表示：

$$L_{SA} = L_{dis} + L_{con}, \tag{4}$$

其中 L_{dis} 通过交叉熵损失强制身份一致性，L_{con} 通过原始特征和攻击特征之间的距离强制特征一致性。

3.2. 流 2：基于记忆的跨视图适应

从空中视角和地面视角之间所产生的显著几何和透视变化是跨平台重识别中最具挑战性的一部分。我们的第二个流引入了一种新颖的基于记忆的架构，通过量身定制的表示学习和自适应的跨视角对齐，明确地建模并弥合这些视角差异。我们创新的核心是开发了专门的视角特定记忆表示 [40, 44, 1] ，这些表示分别

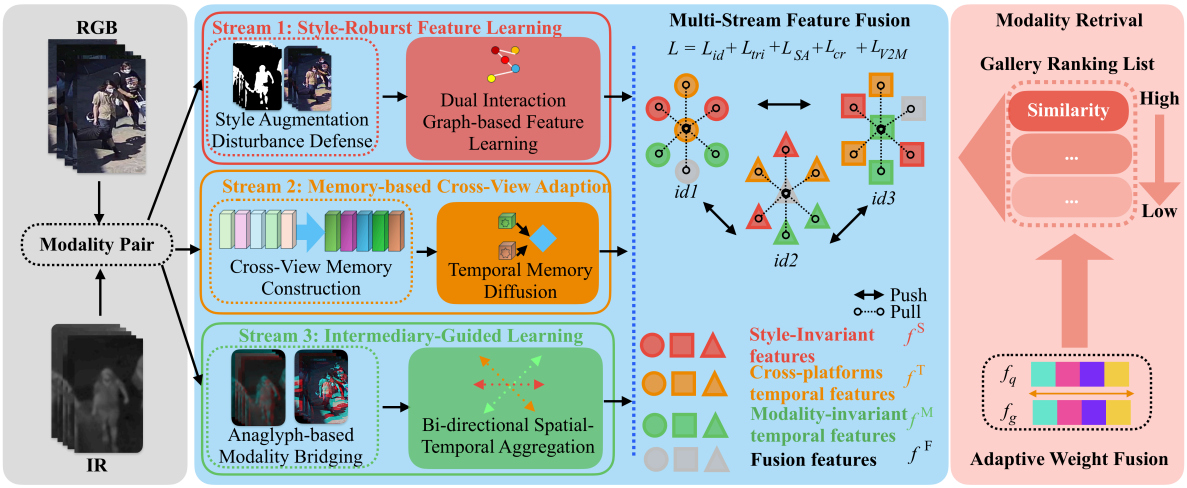

Figure 3. 我们提出的三流架构概述：(1) 风格鲁棒流，通过风格增强和图学习应对外观变化；(2) 基于记忆的跨视角流，用于连接航拍与地面视角；(3) 中介引导的时间流，用于解决 RGB-IR 模态间的差距。这些流在一个融合模块中结合，实现高效的跨平台、跨模态行人再识别。

对空中和地面视角进行建模：

$$M_{y_i}^{aerial} = \frac{1}{N_i^{aerial}} \sum_{v_a \in y_i^{aerial}} v_a, \tag{5}$$

$$M_{y_i}^{ground} = \frac{1}{N_i^{ground}} \sum_{v_a \in y_i^{ground}} v_a, \tag{6}$$

其中 $M_{y_i}^{aerial}$ 和 $M_{y_i}^{ground}$ 分别是身份 y_i 从空中和地面视角获得的记忆表示，v_a 表示从视觉编码器中提取的序列级特征向量，N_i^{aerial} 和 N_i^{ground} 分别表示身份 y_i 的空中和地面序列的数量。与传统方法使用每个身份单一全局记忆不同，我们的双记忆架构明确捕捉了每个视角的不同特性。我们进一步引入了一种跨平台注意机制，动态更新每种记忆类型：

$$M_{y_i}^{m'} = \mathcal{P}_{y_i}^{m} + M_{y_i}^{m}, \tag{7}$$

其中 $m \in \{aerial, ground\}$ 表示平台类型，$M_{y_i}^{m'}$ 是更新的内存表示，$\mathcal{P}_{y_i}^{m}$ 是通过双分支注意力解码器生成的平台特定提示，该解码器显式地在各个平台之间共享信息，同时保留平台特有特征。我们创新的视频到内存对比损失优化了这些表示：

$$L_{V2M}(y_i) = -\frac{1}{|P(y_i)|} \sum_{p \in P(y_i)} \log \frac{\exp(v_p \cdot M_{y_i}^{m'}/\tau)}{\sum_{j=1}^{B} \exp(v_j \cdot M_{y_i}^{m'}/\tau)}, \tag{8}$$

其中 $P(y_i)$ 表示批次中身份 y_i 的所有正样本，v_p 是正样本的特征向量，v_j 指的是批次中第 j 个样本的特征向量，τ 是温度参数，B 是批次大小。

3.3. 流 3：中介引导的时序学习

可见光和红外模式之间的基本域差距带来了独特的挑战，尤其是在与空地视点变化结合时。我们的第三个流引入了一种新颖的中介引导时间学习模块，专门设计用于弥合这些模式差异，同时保留关键的身份信息。

我们的核心方法通过边缘检测 [12]，生成模态不变的红蓝立体图像表示：

$$a(i,j) = \sum_m \sum_n x(i+m, j+n) A(m,n) + k, \tag{9}$$

，其中 $a(i,j)$ 是立体色差图像 a 的第 (i,j) 个像素，$x(i,j)$ 表示原始图像（RGB 或 IR），$A(m,n)$ 表示具有索引 $m,n \in \{-1,0,1\}$ 的边缘检测算子，k 是一个偏移值。我们的创新在于开发了一种专门的三维交叉注意机制，该机制在加强模态不变信息的同时保留了对空中地面对齐至关重要的时空模式。这一机制由我们的交叉重建约束补充，它显式地最小化模态差距：

$$L_{cr} = \sum_{t=1}^{b \times k} \|a_t^v - R(a_t^{ir})\|_2 + \sum_{t=1}^{b \times k} \|a_t^{ir} - R(a_t^v)\|_2, \tag{10}$$

，其中 a_t^v 和 a_t^{ir} 分别表示来自可见光和红外数据的立体色差特征，R 表示由卷积层组成的重建网络，b 是小批量大小，k 是每个视频片段中的帧数。这种方法迫使模型学习模态不变特征，使其能够应对 RGB 和 IR 数据之间显著的外观差异。

3.4. 多流集成和损失函数

这三个流为跨平台跨模态行人再识别提供了互补的信息。我们通过特征融合模块整合这些流，并使用联合损失函数进行优化：

$$L_{total} = L_{id} + \lambda_1 L_{tri} + \lambda_2 L_{SA} + \lambda_3 L_{cr} + \lambda_4 L_{V2M}, \tag{11}$$

，其中超参数 λ_1 、λ_2 、λ_3 和 λ_4 平衡来自之前描述的三个流的每个损失项的贡献。在推断阶段，融合的特征作为跨域匹配的最终行人表示，根据验证性能对每个流的贡献进行自适应加权。

为了进行全面的评估，我们在两个已建立的可见-红外基准 HITSZ-VCM [16] 和 BUPTCampus [6] 上验证我们的方法。此外，我们还在我们的 AG-VPReID.VIR 数据集上进行了全面的实验。正如表 2 所示，我们严格区分训练和测试身份，保留 326 个 ID（978 个轨迹，24,793 帧）用于训练，剩余的身份用于在四个具有挑战性的评估场景中测试：地面到地面，空中到空中，地面到空中，和空中到地面。每个协议包括可见到红外（V2I，使用可见光图像作为查询，红外图像作为库）和红外到可见（I2V，使用红外图像作为查询，可见光图像作为库）的匹配方向。对于 I2V 实验，我们通过在库集中引入额外的干扰身份进一步增加难度，创造一个更真实的监控场景，其中有 1,184 个干扰 ID，贡献了 30,532 到 34,687 帧，具体取决于平台。我们采用标准的行人再识别评估指标：累积匹配特性（CMC）以及平均平均精度（mAP）[6, 39]，以全面衡量所有排名位置的检索性能。

Table 2. AG-VPReID.VIR 的训练/测试统计数据。Q= 查询，G= 库。V2I: 可见光到红外，I2V: 红外到可见光。

Set	Platforms	Modalities	# IDs	# Tracklets (Q/G)	# Frames (Q/G)
Training	All	–	326	978	24,793
Testing	Ground to Ground	V2I	199	313/199	9,357/3,018
		I2V*	199	199/313	3,018/9,357
	Aerial to Aerial	V2I	174	174/174	6,284/3,115
		I2V*	174	174/174	3,115/6,284
	Ground to Aerial	V2I	116	184/116	5,360/1,701
		I2V*	260	260/260	3,530/10,939
	Aerial to Ground	V2I	260	260/260	10,939/3,530
		I2V*	116	116/184	1,701/5,360

* For all I2V experiments, 1,184 additional distractor IDs are added to the gallery set.

我们的 TCC-VPReID 框架在 NVIDIA A100 GPU 上使用 PyTorch。三种流包括：(1) 使用来自 $U(0.5, 1.5)$ 的风格增强和自适应实例归一化的 ResNet50 的风格鲁棒流；(2) 基于记忆的流，采用 CLIP ViT-B/16 编码器和一个两层的 transformer 解码器；(3) 基于中介指导的流，使用具有双分支架构的左右视觉数据和双向 LSTM。这些骨干网络和网络架构是基于各自论文中性能最好的模型进行选择的。训练使用 Adam 优化器（初始 lr= 3.5×10^{-4} ）并采用余弦退火。每个批次包含 8 个身份 × 4 个序列 × 8 帧。模型训练 120 个周期，使用超参数 $\lambda_1 = 1.0$ ，$\lambda_2 = 1.5$ ，$\lambda_3 = 1.0$ ，和 $\lambda_4 = 1.5$（见附录第 8 节）。我们在 HITSZ-VCM、BUPTCampus 和 AG-VPReID.VIR 数据集上对我们的方法进行了与最新技术对比的评估。为了公平比较，我们专注于地面到地面的评估（表 3 ），在跨平台消融测试中在表 4 和 ?? 中展示。

px

px在现有数据集上的表现。在 HITSZ-VCM 上，我们的方法在 I2V 上达到了 72.16 % /56.43 %（Rank-1/mAP），在 V2I 上达到了 75.92 % /57.84 %，比 SAADG 高出 2.94 % /2.66 % 和 2.79 % /1.75 %。在 BUPTCampus 上，我们在 I2V 上达到 69.73 % /67.25 %，在 V2I 上达到 68.47 % /65.38 %，超出 AuxNet 3.25 % /3.14 % 和 3.24 % /3.19 %。

px

pxAG-VPReID.VIR 上的表现。我们的方法在 AG-VPReID.VIR 上显示出最显著的提升，其中在 I2V 上达到 36.18 % /41.56 %，在 V2I 上达到 46.33 % /59.23 % 。这些相比于先前的最佳方法—SAADG（13.07 % /18.75 % ）、AuxNet（15.70 % /16.46 % ）和 CST（14.23 % /19.37 % ）—有着显著的改进，在 Rank-1 上绝对提升了 20.48-23.11 %，在 I2V 的 mAP 上提升了 22.19-25.10 % 。在 V2I 的表现差距更为显著，我们的方法在 Rank-1 上超越了最佳对手 33.69-35.11 %，在 mAP 上超越了 37.18-39.07 % 。值得注意的是，当转换到 AG-VPReID.VIR 的跨平台设置时，所有方法的性能都经历了显著的下降。虽然 SAADG 的 Rank-1 准确率从 69.22 % 急剧下降到 13.07 %（56.15 % 的相对下降），MITML 从 63.74 % 下降到 12.16 %（51.58 % 的下降），但我们的方法在各数据集上维持了更高的性能。

3.5. 消融研究

我们分析了 TCC-VPReID 框架中各个组件的贡献。流被标记为 St1（风格鲁棒特征学习）、St2（基于记忆的跨视角适应）和 St3（中介引导的时间学习）。结果在表格 4 和表格 ?? 中。

个别流性能。St2 一直优于其他单独的流（航拍 → 地面 I2V 的排名为 11.21 %，V2I 为 40.38 %），显示了时间记忆对跨域挑战的有效性。St1 表现次佳，突显了风格不变特征的重要性，而 St3 个人表现最低。

px

组合流。St12 组合相比于单独的流显著提高了性能（航拍 → 地面 I2V 的 Rank-1 为 15.52 %，而 St2 单独为 11.21 %）。完整的三流模型 (St123) 实现了最佳的整体结果，确认了每个流的互补贡献。

px

模态方向的不对称性。性能在 V2I 和 I2V 方向之间显著不同，Aerial → 和 Ground 分别显示 46.54 % 和 19.83 % 的 Rank-1。这种不对称性表明，在搜索可见图库时，红外查询面临更大挑战，可能是由于热成像中的信息损失。

px

平台比较。同平台的跨模态检索（对空中 → 航空 RGB 和 → 红外的 Rank-1 为 33.91 %）优于跨平台场景（对空中 → 地面 I2V 的结果为 19.83 %），这量化了跨平台变化超越模态差异带来的挑战。同样的观察可以应用于地面 → 地面。附加的跨平台结果详见表 5 。

3.6. 定性结果

图 4 比较了我们的 TCC-VPReID 方法与基线方法 [45] 在跨平台跨模态设置下的表现。结果显示在所有场景中排名显著改善（例如从航拍-红外到地面-RGB 的排名从第 16 上升到第 1），证实了我们的三流架构有效解决了视角和模态的变化。

Table 3. 跨模态人员重新识别方法的综合比较。

Method	Venue	Ground ↔ Ground								Ground ↔ Ground			
		HITSZ-VCM				BUPTCampus				AG-VPReID.VIR			
		I2V		V2I		I2V		V2I		I2V		V2I	
		R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP
Image-based Methods													
AlignGAN [25]	ICCV'19	42.15	27.43	44.62	29.73	27.99	30.32	35.37	35.13	4.12	6.32	3.94	8.21
LbA [23]	ICCV'21	46.38	30.69	49.30	32.38	32.09	32.93	39.07	37.06	5.23	7.14	4.82	9.43
MPANet [30]	CVPR'21	46.51	35.26	50.32	37.80	33.45	34.17	40.56	38.22	5.94	8.21	5.63	10.55
SPOT [2]	TIP'22	52.23	37.10	53.67	38.96	37.52	38.24	44.12	41.03	7.32	9.87	6.42	11.63
HCT [17]	TMM'20	55.39	38.62	57.25	39.73	38.92	39.47	45.62	42.18	7.83	10.54	7.21	12.37
DDAG [37]	ECCV'20	54.62	39.26	59.03	41.50	40.44	40.86	46.30	43.05	8.12	11.25	7.83	13.42
MMN [43]	MM'21	53.24	39.82	56.43	41.47	40.86	41.71	43.70	42.80	7.95	11.08	7.42	13.15
CAJL [36]	ICCV'21	56.59	41.49	60.13	42.81	40.49	41.46	45.00	43.61	9.12	12.86	8.76	15.32
VSD [24]	CVPR'21	54.53	41.18	57.52	43.45	41.27	42.05	47.84	44.63	8.45	12.03	8.12	14.27
AGW [38]	TPAMI'21	47.23	36.12	51.85	38.50	36.38	37.36	43.70	41.10	6.74	9.23	6.17	11.08
DEEN [42]	CVPR'23	65.32	50.16	68.42	50.83	49.81	48.59	53.70	50.43	11.23	16.45	10.54	19.23
SGIEL [8]	CVPR'23	67.65	52.30	70.23	52.54	48.32	47.15	51.42	49.27	12.43	17.92	11.21	20.45
Video-based Methods													
TCLNet [11]	ECCV'20	48.32	36.45	52.13	38.52	37.15	36.87	41.32	38.92	9.23	12.54	8.37	15.76
CAViT [29]	ECCV'22	54.83	40.26	58.18	41.63	42.63	41.87	46.52	43.15	10.45	14.83	9.72	17.63
DART [33]	CVPR'22	62.85	44.26	63.92	46.43	52.43	49.10	53.33	50.45	11.56	15.82	10.23	18.76
MITML [16]	CVPR'22	63.74	45.31	64.54	47.69	49.07	47.50	50.19	46.28	12.16	16.93	10.87	19.42
AuxNet [6]	TIFS'23	51.05	45.99	54.58	48.70	66.48	64.11	65.23	62.19	15.70	16.46	12.89	22.57
CST[9]	TMM'24	69.44	51.16	72.64	51.16	54.92	52.87	56.35	53.26	14.23	19.37	12.64	22.05
IBAN [12]	TCSVT'23	65.03	48.77	69.58	50.96	53.42	51.36	54.17	52.43	6.03	9.46	3.83	10.16
SAADG [45]	ACM MM'23	69.22	53.77	73.13	56.09	55.76	53.42	57.83	54.82	13.07	18.75	11.82	21.14
Ours	-	72.16	56.43	75.92	57.84	69.73	67.25	68.47	65.38	36.18	41.56	46.33	59.23

Table 4. 跨平台跨模态 AG-VPReID.VIR 数据集的综合结果。

Method	Aerial → Ground										Ground → Aerial									
	I2V					V2I					I2V					V2I				
	R1	R5	R10	R20	mAP	R1	R5	R10	R20	mAP	R1	R5	R10	R20	mAP	R1	R5	R10	R20	mAP
St1	7.76	17.24	22.41	35.34	9.54	7.31	23.46	39.62	52.31	17.46	12.31	26.54	40.00	47.69	20.64	14.13	27.72	44.02	57.61	23.05
St2	11.21	28.45	36.21	45.69	17.68	40.38	70.77	82.69	92.31	54.40	27.31	52.69	63.08	71.92	39.27	27.17	50.54	65.76	79.35	39.25
St3	2.59	11.21	17.24	22.41	4.86	5.38	13.46	22.69	36.92	11.50	3.08	14.23	18.46	26.92	9.10	7.07	20.65	33.15	47.28	15.61
St12	15.52	31.90	43.10	51.72	20.91	46.92	77.31	86.15	92.31	59.71	34.23	56.54	70.77	78.85	45.62	30.43	56.52	67.39	80.98	42.45
St13	9.48	18.97	28.45	37.93	10.94	8.08	26.15	41.92	56.92	17.96	13.85	28.46	38.08	48.46	21.43	14.67	32.07	45.65	59.24	24.11
St23	18.10	34.48	42.24	45.69	21.22	41.54	74.23	85.00	92.31	56.31	31.92	55.38	66.54	78.08	43.01	31.52	53.80	66.85	78.80	42.49
St123	19.83	31.90	43.10	51.72	22.61	46.54	76.54	85.38	91.92	59.69	34.23	56.54	70.77	78.85	45.62	31.52	57.07	67.39	80.98	42.92

Table 5. AG-VPReID.VIR 地面对地面与空中对地面场景（红外对可见光）的性能比较。

Method	Ground → Ground		Aerial → Ground	
	R1	mAP	R1	mAP
MITML [CVPR'22]	12.16	16.93	3.95	6.45
SAADG [ACM MM'23]	13.07	18.75	4.21	7.83
CST [TMM'24]	14.23	19.37	4.87	8.12
Ours	36.18	41.56	19.83	22.61

4. 结论

我们介绍了 AG-VPReID.VIR，这是第一个用于视频基础上的跨模态人员再识别的全面数据集，涵盖航拍和地面平台，包含来自 1,837 个身份的 4,861 个轨迹。我们提出的 TCC-VPReID 框架通过风格稳健的特征学习、基于记忆的跨视图适应和中介引导的时间学习，有效解决了视角变化和模态差异。实验结果表明 TCC-VPReID 在 Rank-1 准确率和 mAP 上均超过现有方法 20 %，以此为强健的跨平台监控系统设立了新的挑战性基准。

References

[1] G. C. Bertocco, F. Andaló, and A. Rocha. Unsupervised and self-adaptative techniques for cross-domain person re-identification. IEEE Transactions on Information Forensics and Security, 16:4419–4434, 2021.

[2] C. Chen, M. Ye, M. Qi, J. Wu, J. Jiang, and C.-W. Lin. Structure-aware positional transformer for visible-infrared person re-identification. IEEE Transactions on Image Processing, 31:2352–2364, 2022.

[3] D. Chung, K. Tahboub, and E. J. Delp. A two stream siamese convolutional neural network for person re-identification. In ICCV, pages 1983–1991, 2017.

[4] P. Dai, R. Ji, H. Wang, Q. Wu, and Y. Huang. Cross-modality person re-identification with generative adversarial training. In IJCAI, volume 1, page 6, 2018.

[5] M. Dreuw and ORB-HD. deface: Video anonymization by face detection, 2023. Python package version 1.5.0.

[6] Y. Du, C. Lei, Z. Zhao, Y. Dong, and F. Su. Video-based visible-infrared person re-identification with auxiliary samples. IEEE Transactions on Information Forensics and Security, 19:1313–1325, 2023.

Figure 4. 在 AG-VPReID.VIR 数据集下，对于 I2V 和 V2I 查询（I：红外，V：可见光）进行航拍-地面跨平台设置的前 20 名排名结果的可视化。红色边框表示错误匹配，绿色表示正确匹配。

[7] C. Eom, G. Lee, J. Lee, and B. Ham. Video-based person re-identification with spatial and temporal memory networks. In ICCV, pages 12036–12045, 2021.

[8] J. Feng, A. Wu, and W.-S. Zheng. Shape-erased feature learning for visible-infrared person re-identification. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 22752–22761, 2023.

[9] Y. Feng, F. Chen, J. Yu, Y. Ji, F. Wu, T. Liu, S. Liu, X.-Y. Jing, and J. Luo. Cross-modality spatial-temporal transformer for video-based visible-infrared person re-identification. IEEE Transactions on Multimedia, 26:6582–6594, 2024.

[10] X. Gu, H. Chang, B. Ma, H. Zhang, and X. Chen. Appearance-preserving 3d convolution for video-based person re-identification. In ECCV, pages 228–243. Springer, 2020.

[11] R. Hou, H. Chang, B. Ma, S. Shan, and X. Chen. Temporal complementary learning for video person re-identification. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16, pages 388–405. Springer, 2020.

[12] H. Li, M. Liu, Z. Hu, F. Nie, and Z. Yu. Intermediary-guided bidirectional spatial–temporal aggregation network for video-based visible-infrared person re-identification. IEEE Transactions on Circuits and Systems for Video Technology, 33(9):4962–4972, 2023.

[13] J. Li, S. Zhang, and T. Huang. Multi-scale 3d convolution network for video based person re-identification. In AAAI, volume 33, pages 8618–8625, 2019.

[14] Z. Li, W. Liu, X. Chang, L. Yao, M. Prakash, and H. Zhang. Domain-aware unsupervised cross-dataset person re-identification. In Advanced Data Mining and Applications, 2019.

[15] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In CVPR, pages 2197–2206, 2015.

[16] X. Lin, J. Li, Z. Ma, H. Li, S. Li, K. Xu, G. Lu, and D. Zhang. Learning modal-invariant and temporal-memory for video-based visible-infrared person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 20973–20982, 2022.

[17] H. Liu, X. Tan, and X. Zhou. Parameter sharing exploration and hetero-center triplet loss for visible-thermal person re-identification. IEEE Transactions on Multimedia, 23:4414–4425, 2020.

[18] W. Liu, X. Chang, L. Chen, D. Q. Phung, X. Zhang, Y. Yang, and A. G. Hauptmann. Pair-based uncertainty and diversity promoting early active learning for

person re-identification. ACM Transactions on Intelligent Systems and Technology (TIST), 11:1–15, 2020.

[19] Y. Liu, Z. Yuan, W. Zhou, and H. Li. Spatial and temporal mutual promotion for video-based person re-identification. In AAAI, volume 33, pages 8786–8793, 2019.

[20] N. McLaughlin, J. M. Del Rincon, and P. Miller. Recurrent convolutional network for video-based person re-identification. In CVPR, pages 1325–1334, 2016.

[21] D. T. Nguyen, H. G. Hong, K. W. Kim, and K. R. Park. Person recognition system based on a combination of body images from visible light and thermal cameras. Sensors, 17(3):605, 2017.

[22] H. Nguyen, K. Nguyen, S. Sridharan, and C. Fookes. Ag-reid. v2: Bridging aerial and ground views for person re-identification. IEEE Transactions on Information Forensics and Security, 2024.

[23] H. Park, S. Lee, J. Lee, and B. Ham. Learning by aligning: Visible-infrared person re-identification using cross-modal correspondences. In Proceedings of the IEEE/CVF international conference on computer vision, pages 12046–12055, 2021.

[24] X. Tian, Z. Zhang, S. Lin, Y. Qu, Y. Xie, and L. Ma. Farewell to mutual information: Variational distillation for cross-modal person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1522–1531, 2021.

[25] G. Wang, T. Zhang, J. Cheng, S. Liu, Y. Yang, and Z. Hou. Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment. In ICCV, pages 3623–3632, 2019.

[26] Z. Wang, Z. Wang, Y. Wu, J. Wang, and S. Satoh. Beyond intra-modality discrepancy: A comprehensive survey of heterogeneous person re-identification. In International Joint Conference on Artificial Intelligence, 2020.

[27] Z. Wang, Z. Wang, Y. Zheng, Y.-Y. Chuang, and S. Satoh. Learning to reduce dual-level discrepancy for infrared-visible person re-identification. In CVPR, pages 618–626, 2019.

[28] A. Wu, W.-S. Zheng, H.-X. Yu, S. Gong, and J. Lai. Rgb-infrared cross-modality person re-identification. In Proceedings of the IEEE international conference on computer vision, pages 5380–5389, 2017.

[29] J. Wu, L. He, W. Liu, Y. Yang, Z. Lei, T. Mei, and S. Z. Li. Cavit: Contextual alignment vision transformer for video object re-identification. In European Conference on Computer Vision, pages 549–566. Springer, 2022.

[30] Q. Wu, P. Dai, J. Chen, C.-W. Lin, Y. Wu, F. Huang, B. Zhong, and R. Ji. Discover cross-modality nuances for visible-infrared person re-identification. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 4330–4339, 2021.

[31] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Ouyang, and Y. Yang. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 5177–5186, 2018.

[32] S. Xu, Y. Cheng, K. Gu, Y. Yang, S. Chang, and P. Zhou. Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In ICCV, pages 4733–4742, 2017.

[33] M. Yang, Z. Huang, P. Hu, T. Li, J. Lv, and X. Peng. Learning with twin noisy labels for visible-infrared person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14308–14317, 2022.

[34] M. Ye, X. Lan, Q. Leng, and J. Shen. Cross-modality person re-identification via modality-aware collaborative ensemble learning. IEEE Transactions on Image Processing, 29:9387–9399, 2020.

[35] M. Ye, X. Lan, J. Li, and P. Yuen. Hierarchical discriminative learning for visible thermal person re-identification. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 32, 2018.

[36] M. Ye, W. Ruan, B. Du, and M. Z. Shou. Channel augmented joint learning for visible-infrared recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 13567–13576, 2021.

[37] M. Ye, J. Shen, D. J. Crandall, L. Shao, and J. Luo. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In ECCV, pages 229–247. Springer, 2020.

[38] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi. Deep learning for person re-identification: A survey and outlook. IEEE transactions on pattern analysis and machine intelligence, PP, 2021.

[39] C. Yu, X. Liu, Y. Wang, P. Zhang, and H. Lu. Tf-clip: Learning text-free clip for video-based person re-identification. In Proceedings of the AAAI conference on artificial intelligence, volume 38, pages 6764–6772, 2024.

[40] G. Zhang, H. Zhang, W. Lin, A. K. Chandran, and X. Jing. Camera contrast learning for unsupervised person re-identification. IEEE Transactions on Circuits and Systems for Video Technology, 33(8):4096–4107, 2023.

[41] S. Zhang, W. Luo, D. Cheng, Q. Yang, L. Ran, Y. Xing, and Y. Zhang. Cross-platform video person reid: A new benchmark dataset and adaptation approach. In European Conference on Computer Vision (ECCV), 2024.

[42] Y. Zhang and H. Wang. Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification. arXiv preprint arXiv:2303.14481, 2023.

[43] Y. Zhang, Y. Yan, Y. Lu, and H. Wang. Towards a unified middle modality learning for visible-infrared person re-identification. In Proceedings of the 29th ACM International Conference on Multimedia, pages 788–796, 2021.

[44] D. Zheng, J. Xiao, M. Sun, H. Bai, and J. Hou. Plausible proxy mining with credibility for unsupervised person re-identification. IEEE Transactions on Circuits and Systems for Video Technology, 33(7):3308–3318, 2022.

[45] C. Zhou, J. Li, H. Li, G. Lu, Y. Xu, and M. Zhang. Video-based visible-infrared person re-identification via style disturbance defense and dual interaction. In Proceedings of the 31st ACM International Conference on Multimedia, pages 46–55, 2023.

AG-VPReID.VIR: 为
基于视频的可见光-红外人员重识别搭建空中与地面平台的桥梁

Supplementary Material

5. 数据集收集

图 5 展示了我们用于 AG-VPReID.VIR 数据集收集的战略性摄像头布置，这些摄像头分布在大学校园内。地图突出显示了用于获取跨平台和跨模态数据的多样化感知基础设施：仅可见光相机（V）、仅红外相机（IR）以及同时捕捉可见光和红外模态的双模态相机（VIR）。我们策略性地布置这些摄像头以最大化覆盖范围，同时确保不同感知模态和视角之间的最小重叠。无人机被安置在固定的悬停位置（在地图上标记出），但在不同的高度（15 米、25 米和 45 米）以捕获多样化的空中视角，而地面 CCTV 和可穿戴摄像机提供了互补的视点。这一全面的感知网络使得在不同环境条件下收集同步多视角数据，形成了具有变化照明、遮挡和视角的具有挑战性的场景，逼真地反映了实际的监控情境。

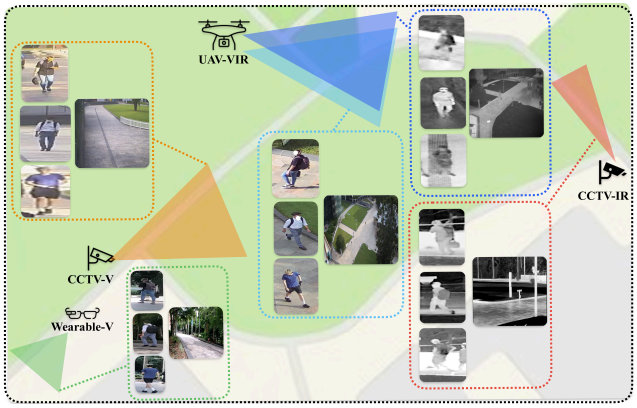

Figure 5. 数据采集图。V 代表可见光，IR 表示红外，VIR 代表同时使用可见光和红外相机。

6. 数据集分布分析

图 6 展示了我们的 AG-VPReID.VIR 数据集中不同相机类型和模式的数据分布。分布显示了在 RGB 和 IR 模式之间的刻意平衡，同时在平台之间保持多样性。无人机 RGB 占数据集的最大部分 (29.2 %)，提供了丰富的航拍可见光视角，而闭路电视 IR 占第二大份额 (25.6 %)，提供了大量地面热成像。闭路电视 RGB (18.9 %)、无人机 IR (11.5 %) 和可穿戴设备 RGB (14.8 %) 的互补特性确保了在所有部署场景中的全面覆盖。这种平衡分布对于训练稳健的跨模式跨平台再识别模型至关重要，因为它为算法提供了足够的例子来学习每种视角-模式组合的独特特征，特别是我们数据集中独有的具有挑战性的航拍红外视角。

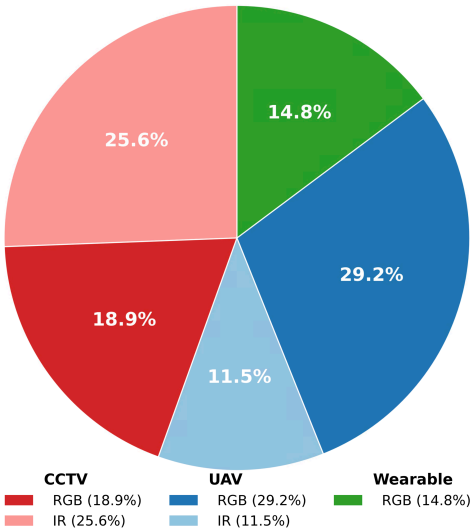

Figure 6. 我们 AG-VPReID.VIR 数据集中不同相机类型和模式（RGB 和 IR）下身份存在的分布。

7. 所提出架构中流的挑战和关键组件

表 6 展示了我们 TCC-VPReID 框架的完整结构，详细说明了每个流如何解决空地可见光-红外人重识别中的特定挑战。这三个互补的流协同工作以克服跨平台跨模态场景中遇到的复杂变化。流 1 专注于风格稳健特征学习，通过风格增强、跨视图图交互和领域对抗匹配等组件来处理平台间/模态内的变化、帧级风格变化和外观畸变。流 2 通过基于记忆的适应性解决时间和跨视图挑战，构建维护空中和地面视角身份一致性的特定平台记忆。流 3 针对 RGB 和 IR 图像之间的基本模态差距，使用立体图像中间学习和双向聚合。这些流通过特征融合模块集成，以保留互补信息，同时调和跨平台和模态的矛盾。这种多流方法使我们的框架能够有效地解开表征空地可见光-红外人重识别中的外观、视角和模态变化。

8. 超参数消融研究

我们进行了消融研究，以确定损失函数超参数（λ_1，λ_2，λ_3 和 λ_4）的最优值。表 7 展示了我们模型在不同超参数配置下在 Aerial→Ground I2V 协议上的性能，该协议是我们数据集中最具挑战性的场景之一。

如表 7 所示，我们发现设置 $\lambda_1 = 1.0$、$\lambda_2 = 1.5$、$\lambda_3 = 1.0$ 和 $\lambda_4 = 1.5$ 在我们最具挑战性的检索场景中达到了最佳性能。结果表明：

- 三元组损失（λ_1）在权重为 1.0 时表现最佳，这在平衡其与其他损失的贡献方面取得了最好效果。

Table 6. TCC-VPReID 概述：一种用于跨平台跨模态基于视频的人体重识别的三流架构

Stream	Challenges Addressed	Key Components
Stream 1: Style-Robust Feature Learning	• Intra/inter-modal variations • Frame-level style variations • Appearance distortions across platforms	• Style Augmentation & Disturbance Defense • Cross-View & Cross-Modal Graph Interaction • Domain-Adversarial Alignment
Stream 2: Temporal Memory Adaptation	• Aerial-ground viewpoint variations • Temporal misalignment in sequences • Unstable motion patterns across platforms	• Cross-View Memory Construction • Transformer-Based Video Encoder • Temporal Memory Diffusion
Stream 3: Modality-Invariant Representation	• Modality gap (visible vs. infrared) • Noisy frames in aerial/ground videos • Information asymmetry between modalities	• Anaglyph-based Modality Bridging • Bidirectional Spatial-Temporal Aggregation • Modality Decoupling
Feature Fusion	• Preserving complementary information • Reconciling cross-platform/modality features • Balancing spatial, temporal, and modality cues	• Multi-level Feature Integration • Cross-stream Knowledge Transfer • Joint Optimization Framework

Table 7. 使用 Aerial→Ground I2V 性能对超参数 λ_1 ， λ_2 ， λ_3 和 λ_4 进行消融研究。

λ_1	λ_2	λ_3	λ_4	Rank-1	mAP
0.5	1.0	0.5	1.0	15.21	18.43
0.5	1.0	1.0	1.5	17.46	20.12
0.5	1.5	1.0	1.5	18.32	21.05
1.0	1.0	1.0	1.0	17.65	20.33
1.0	1.5	0.5	1.5	18.75	21.42
1.0	1.5	1.0	1.5	19.83	22.61
1.0	2.0	1.0	1.5	19.12	21.98
1.5	1.5	1.0	1.5	18.96	21.75
1.0	1.5	1.5	1.5	18.54	21.30
1.0	1.5	1.0	2.0	19.07	21.83

- 风格攻击损失（ λ_2 ）由于具有更高的权重 1.5 而受益，强调风格不变特征对于跨平台场景的重要性。
- 交叉重建损失（λ_3 ）在权重为 1.0 的中等水平时最为有效，帮助弥合模态差距而不压制其他组件。
- 视频到记忆的对比损失（ λ_4 ）在 1.5 时表现最佳，表明了时间记忆对于跨视角适应的重要性。

我们观察到，将 λ_2 和 λ_4 增加到超过 1.5 会导致性能下降，这表明过度强调这些组件可能会破坏不同学习目标之间的平衡。