WHU-MARS: A Multispectral Aerial-Ground Benchmark Towards Any-Scenario Person Re-Identification

Supplementary Material

This supplementary material provides additional details and analysis in two parts.

- **WHU-MARS Dataset Details**: additional statistics, collection and annotation pipeline, dataset versions and splits, and privacy and ethical considerations.
- **Discussion of Existing Datasets and Methods**: extended analysis of why current datasets and unified methods do not directly meet the requirements of AS-ReID.

We also briefly discuss the limitations of WHU-MARS and potential future extensions.

1. WHU-MARS Dataset Details

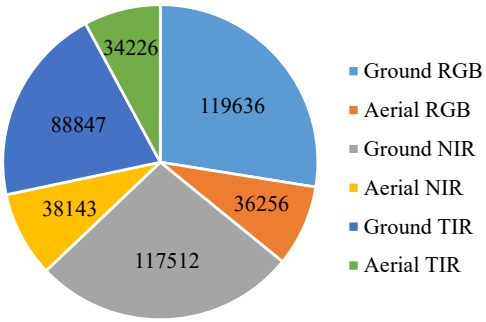
1.1. Dataset Statistics

To further characterize WHU-MARS and to illustrate its suitability for AS-ReID, we provide additional visual statistics in Figure 1. Unless otherwise specified, all statistics are reported for the full WHU-MARS-2337 version.

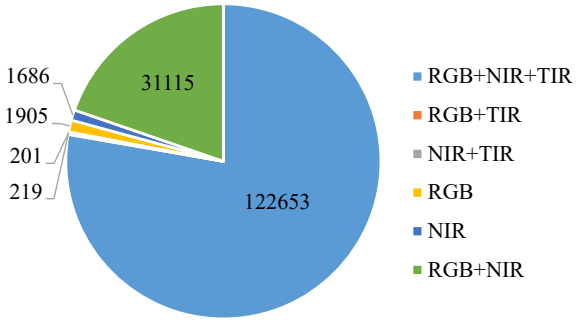
As shown in Figure 1(a), the three modalities contain comparable numbers of images. Both ground and aerial platforms employ tri-spectral RGB-NIR-TIR sensors that record all streams synchronously, leading to similar data volumes per modality. This design helps mitigate overfitting to a dominant spectrum and encourages learning modality-invariant representations. Each identity is observed in at least two camera views, which is consistent with the cross-camera retrieval objective in ReID.

Figure 1(b) summarizes the composition of frame-synchronized multi-modal paired samples, where a sample is considered paired when images of the same identity are captured at the same time frame across multiple modalities. Approximately 77.7% of samples form complete RGB-NIR-TIR triplets. Due to the different aspect ratios and fields-of-view (FOVs) of the TIR cameras, about 19.7% of samples contain only RGB-NIR pairs, while occasional occlusions or missed detections account for the remaining 2.5% with other partial modality combinations. This mixture of complete and incomplete modality combinations reflects realistic deployment constraints while still providing abundant frame-synchronized tri-spectral triplets for unified benchmarking under multiple protocols.

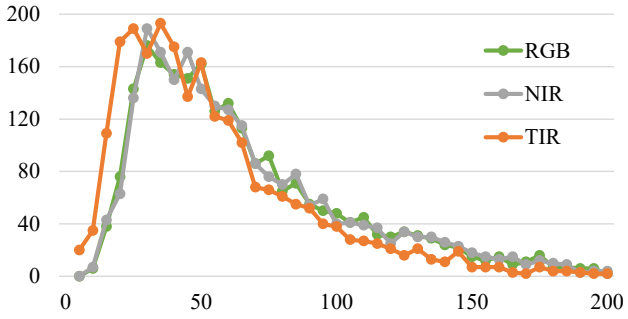
Figure 1(c) reports the distribution of identities across different images. On average, each identity has more than 50 images per modality, and 98% of identities have at least 40 images. Such dense coverage across modalities and views introduces substantial intra-identity appearance variation and supports robust training and evaluation.



(a) The number of images of each modality and viewpoint.



(b) The number of each multi-modal paired sample type.



(c) The distribution of identities across different images.

Figure 1. Statistics of the WHU-MARS-2337 dataset.

1.2. Data Collection

To reflect realistic multi-scenario deployments, WHU-MARS is collected on a university campus using two types of tri-spectral camera systems at diverse viewpoints.

(A) Custom tri-spectral PTZ camera. For ground-view capture, we primarily deploy a custom tri-spectral pan-tilt-zoom (PTZ) system that integrates two 1920×1080 cameras for RGB and NIR imaging and one 640×480 TIR camera.

The three sensors are mounted in a PTZ housing and record synchronously at 30 fps.

(B) DJI H20T gimbal camera. For aerial-view capture, we adopt the DJI Zenmuse H20T as the tri-spectral imaging platform. It contains a 3840×2160 main camera that supports RGB/NIR switching, a 1920×1080 wide-angle RGB camera, and a 640×512 TIR camera, all mounted on the same stabilized gimbal and recording synchronously at 25 fps. In our setup, the main camera operates in NIR mode, the wide-angle camera is used as the RGB channel, and the TIR camera provides the TIR channel.

Ground cameras 1, 2, 4, and 5 use the custom PTZ system (type A), whereas ground camera 3 and aerial cameras 6 and 7 use the DJI H20T platform (type B). The use of two hardware platforms naturally introduces intra-modality variation in resolution, FoVs, and lens characteristics, making WHU-MARS more representative of heterogeneous real-world deployments.

Ground cameras are installed at a height of approximately 1.5m, capturing individuals from near eye level. UAVs operate at altitudes between 20-50m, covering large areas from oblique viewpoints. Data collection spans multiple functional areas on campus, including teaching buildings, dormitories, canteens, supermarkets, and bus stops. This variety of environments introduces substantial variations in clothing styles, backgrounds, crowd density, and motion patterns. Data collection is conducted from July to the following January between 09:00 and 21:00 local time, covering 13 sessions in total and yielding more than 38 hours of raw video. This schedule covers various illuminations, multiple seasons, and diverse weather, providing a broad range of real-world scenarios.

1.3. Data Annotation

Compared with conventional single-modality, single-view ReID datasets, the heterogeneous multi-modal and multi-view scenarios in WHU-MARS pose substantially greater challenges for annotation. To address these challenges, we design a unified annotation pipeline that combines RGB-based tracking, cross-camera identity association, and cross-modality propagation. The pipeline relies on automatic tools and extensive human verification to produce reliable, high-quality annotations for WHU-MARS.

Most existing multi-object trackers are designed for RGB videos, degrade significantly in NIR, and become highly unreliable in TIR due to low contrast and substantially different appearance characteristics. Moreover, independently tracking each modality would lead to inconsistent identity labels across modalities, which is particularly problematic for MM-ReID, where frame-synchronized tri-spectral triplets are required.

To avoid these issues, we first apply ByteTrack [9] on the RGB stream of each camera to generate initial bounding boxes and tracklets. Even on RGB videos, nighttime and low-illumination conditions remain challenging, so human annotators carefully review the tracking results to correct inaccurate bounding boxes and identity assignments. Annotators also watch the videos and use a pre-trained ReID model as an auxiliary tool to assist cross-camera identity association. Following standard ReID practice, we retain only identities that appear in at least two camera views so that each identity contributes to cross-camera retrieval.

Once the RGB annotations are finalized, we propagate them to the other modalities to build tri-spectral annotations. For ground cameras using the custom PTZ system (type A), the RGB and NIR cameras share nearly identical FoVs. We therefore perform frame-level cross-modal registration and automatically transfer RGB bounding boxes to the corresponding NIR streams. By contrast, for the TIR stream of the custom PTZ system (type A) and for the H20T platform (type B), differences in FoV and lens configuration lead to more pronounced cross-modal misalignment. In these cases, annotators manually refine and verify the bounding boxes in the NIR and TIR images with reference to the synchronized RGB frames.

This pipeline yields tri-spectral aligned bounding boxes with consistent identity labels across views and modalities. The combination of RGB-based tracking, cross-camera association, cross-modality propagation, and systematic manual verification ensures the reliability and quality of the final WHU-MARS annotations.

1.4. Dataset Versions

As discussed in Section 3.3 of the main paper, we provide two official versions of WHU-MARS: WHU-MARS-2337 and WHU-MARS-1000. WHU-MARS-1000 retains only frame-synchronized tri-spectral triplets and is designed for unified benchmarking of AG-ReID, VI-ReID, MM-ReID, and AS-ReID at a moderate scale. In contrast, WHU-MARS-2337 retains all available images and identities, making it more representative of real-world deployments with unpaired modalities and suitable for more challenging AS-ReID evaluation and multispectral pretraining.

When constructing these two versions, we follow three design principles: (i) the training sets of both versions should contain as many cross-view multispectral observations as possible to facilitate the learning of modality-invariant representations; (ii) the training and test identities of WHU-MARS-1000 should be subsets of the corresponding splits in WHU-MARS-2337, enabling fair comparison across different data scales; (iii) the identities selected for WHU-MARS-1000 should contain sufficient frame-synchronized tri-spectral triplets to support unified evaluation under AG-, VI-, MM-, and AS-ReID protocols on the same identity set. Based on these principles, we first select a high-quality synchronized tri-spectral subset of 1,500 iden-

tities from the full WHU-MARS collection. For each identity in this subset, we require at least 8 frame-synchronized tri-spectral triplets in every available camera view. This constraint ensures rich variation in appearance within and across views while providing sufficient tri-spectral triplets to support MM-ReID.

We then construct WHU-MARS-1000 by randomly sampling 1,000 identities from this 1,500-identity subset. These 1,000 identities are evenly split into 500 training identities and 500 test identities, and we only retain their frame-synchronized tri-spectral triplets in this version. We further construct WHU-MARS-2337 as a larger and more heterogeneous version. Specifically, the training set of WHU-MARS-2337 consists of the 500 training identities from WHU-MARS-1000 together with the remaining 500 identities from the 1500-identity tri-spectral subset, resulting in 1,000 training identities in total. The test set of WHU-MARS-2337 contains the remaining 1,337 identities, including the 500 test identities used in WHU-MARS-1000 and an additional 837 identities from the rest of the dataset. For WHU-MARS-2337, we retain all images of these identities, including both fully paired and incomplete multi-modal samples. Although the test set of WHU-MARS-2337 includes identities with incomplete modality combinations, each identity still has observations in multiple modalities and camera views, preserving the heterogeneous multi-scenario setting required for AS-ReID.

In summary, the two versions are complementary by design. Both training sets contain a large number of cross-view tri-spectral identities, encouraging models to learn unified, modality-irrelevant representations under a consistent training setting. At the same time, the WHU-MARS-2337 test set introduces many identities with incomplete modality combinations and unbalanced viewpoints, making AS-ReID evaluation closer to real deployment scenarios, while WHU-MARS-1000 provides a moderate-scale benchmark for unified multi-protocol comparison.

1.5. Privacy and Ethical Considerations

The construction of WHU-MARS follows strict privacy and ethical guidelines. Data collection was approved by the institutional office, and all recordings were conducted exclusively in publicly accessible campus areas. Cameras were mounted in clearly visible locations, and on-site notices were posted to describe the purpose of data collection and its intended research use. For public release, we provide only cropped person bounding boxes rather than raw videos or full-frame surveillance footage. All visible faces are automatically mosaicked to suppress personally identifiable facial traits. The released dataset is restricted to academic research under a usage agreement that explicitly prohibits any commercial or unethical use. Upon public release, we will also provide a contact channel on the project page so that individuals or institutions can request the removal of specific data if necessary.

2. Discussion of Existing Datasets and Methods

In this section, we discuss how existing multimodal and multi-view ReID datasets and representative unified ReID frameworks relate to the proposed AS-ReID task, with a focus on the mismatch between their underlying assumptions and the any-to-any retrieval setting of AS-ReID.

2.1. Discussion of Existing Datasets

As summarized in Table 1 and Section 2.1 of the main paper, existing multi-modal and multi-view ReID datasets are typically constructed around specific task formulations and predefined scenario pairs. These benchmarks are effective for analyzing specific cross-modality or cross-view conditions. From the AS-ReID perspective, however, identity observations over the full scenario space are often incomplete. Evaluation protocols are therefore usually decomposed into several disjoint sub-tasks, each built on a particular scenario pair. As a result, models are typically trained and evaluated separately for each protocol, rather than learning a single representation that supports any-to-any retrieval across all modalities and viewpoints within one dataset.

Among multi-modal aerial-ground datasets, MP-ReID [3] and AG-VPReID.VIR [8] are closest in spirit to our setting. Both deploy heterogeneous sensors on ground and aerial platforms to capture individuals across RGB and infrared modalities. However, different modalities are typically captured by distinct physical cameras with different placements, so the presence of each identity across modalities and viewpoints is often unbalanced. Many identities appear in only a small subset of scenarios, and comprehensive coverage across all available scenarios is therefore difficult to guarantee. As a result, the official protocols typically partition the raw data into multiple sub-benchmarks tailored to specific tasks (e.g., aerial TIR to ground RGB matching). Correspondingly, specialized models are trained and evaluated separately for each sub-task. This design is well aligned with the original objective of learning specific cross-modality or cross-platform challenges. However, even though these datasets contain multi-modal and multi-view data, identity-level coverage over the scenario space remains fragmented and does not directly support unified any-scenario ReID.

RGBNT201 [10] further enriches multimodal sensing by introducing a paired RGB-NIR-TIR dataset for MM-ReID, with synchronized tri-spectral observations acquired by tri-spectral cameras. However, due to the cost and deployment complexity of such hardware, most training identities in RGBNT201 are observed from only a single camera view, resulting in relatively limited cross-camera variation. During training, models are therefore mainly en-

couraged to exploit complementary information across the three spectra rather than learning camera-invariant representations, which leads to a mismatch with the core goal of cross-camera retrieval in ReID. In addition, RGBNT201 is relatively small in scale, with 201 identities and 14,361 images. This scale is adequate for studying MM-ReID, but is less suitable for training robust AS-ReID models that must generalize across diverse scenarios.

Overall, existing datasets significantly advance their respective tasks. From the AS-ReID perspective, however, their reliance on scene-pair protocols and fragmented identity coverage across scenarios makes them not directly aligned with the requirements of AS-ReID, where each identity should be consistently represented across a broad range of scenarios within a single dataset.

2.2. Discussion of Existing Methods

As discussed in Section 2.2 of the main paper, recent works have explored modeling diverse ReID scenarios within a single network. These frameworks typically target specific modality pairs, collections of tasks, or evaluation protocols, and are highly effective under their intended settings. Nonetheless, their underlying assumptions differ from those of AS-ReID, which requires a unified representation for retrieving the same identity from a heterogeneous gallery spanning all scenarios. Below, we briefly categorize representative methods and clarify these mismatches.

VI-style dual-stream solutions. Visible-Infrared ReID methods focus on bridging a predefined pair of modalities using dual-stream architectures and pairwise alignment losses. Such designs effectively learn modality-invariant features under a predefined two-modality setting, but they usually rely on modality-specific branches and pairwise objectives tailored to particular modality pairs. When extended to richer scenarios, the number of branches and pairwise objectives grows quickly, making these architectures difficult to scale to the broader scenario space of AS-ReID.

Prompt-based multi-task systems. Prompt-driven frameworks such as Uni-Prompt [3], Instruct-ReID [4], and VersReID [11] aim to unify multiple ReID tasks within one network by conditioning on task labels or textual instructions. Different tasks (e.g., Tr-, VI-, clothes-changing, occlusion ReID) are often defined on heterogeneous datasets whose identity sets are typically disjoint. As a result, these models are optimized to handle multiple task-specific retrieval settings given a prompt, and evaluation is typically performed separately for each task rather than on a single mixed, heterogeneous gallery. This task-centric formulation does not explicitly enforce a unified identity representation that consistently describes the same person across diverse scenarios within one dataset, as required by AS-ReID.

Unified-representation pipelines. Unified-representation pipelines such as AIO [5] and ReID5o [12] further advance modality-invariant ReID by mapping heterogeneous inputs into a shared feature space via modality-specific tokenizers or multi-expert routing. Conceptually, these methods are the closest to the goal of learning unified representations. However, their retrieval formulations are typically asymmetric: RGB is often treated as the dominant modality, and retrieval is mainly defined with RGB galleries as the target, with optimization objectives tailored to this RGB-centered regime. In AS-ReID, by contrast, all modalities and viewpoints are treated symmetrically, and any scenario may act as either query or gallery, which goes beyond the original design target of these pipelines.

Protocol-extension approaches. A complementary line of work extends existing benchmarks with additional protocols while keeping the underlying data and task definitions fixed. Uni-AT [6] targets anytime retrieval by introducing scenario-specific class tokens and mixture-of-experts heads for different conditions. Building on VI-ReID, MixReID [7] and MixER [1] jointly learn intra-modality and cross-modality consistency, but they employ distinct feature spaces or retrieval rules for different query types. Meanwhile, MM-ReID methods such as MDReID [2] address modality mismatch under tri-spectral pairing assumptions. These approaches are effective in their respective regimes, but they typically introduce protocol-specific heads, branches, or retrieval strategies, aiming to support multiple protocols within a single model rather than learning a unified representation as required by AS-ReID.

In summary, existing unified methods advance ReID toward greater generality from different angles, including cross-modality alignment, prompt-based multi-tasking, multi-modal unified encoding, and multi-protocol evaluation. However, they are built on assumptions such as predefined modality pairs, task-disjoint identity sets, RGB-centered retrieval, or protocol-specific architectures, which differ from the AS-ReID setting. These differences suggest that adapting existing methods to AS-ReID may require revisiting their retrieval formulations and training objectives, and also highlight the value of a unified training framework designed for any-to-any retrieval within a single model.

3. Limitations and Future Work

While WHU-MARS provides a large-scale multispectral aerial-ground benchmark with diverse scenarios, it currently remains limited to visual modalities. Recent advances in text-based person search (TBPS) and vision-language pre-training indicate that aligning images with language descriptions could further improve the flexibility and generality of ReID systems. A natural direction for future work is to augment WHU-MARS with textual annotations and explore unified training objectives that jointly support visual AS-ReID and TBPS, moving toward a more comprehensive multimodal benchmark and framework.

References

[1] Mahdi Alehdaghi, Rajarshi Bhattacharya, Dai Yannick, Pourya Shamsolmoali, Rafael MO Cruz, and Eric Granger. Mixer: From cross-modal to mixed-modal visible-infrared re-identification. In *WACV*, pages 3431–3440, 2026. 4

[2] Yingying Feng, Jie Li, Jie Hu, Yukang Zhang, Lei Tan, and Jiayi Ji. Mdreid: Modality-decoupled learning for any-to-any multi-modal object re-identification. In *NeurIPS*, 2025. 4

[3] Ruiyang Ha, Songyi Jiang, Bin Li, Bikang Pan, Yihang Zhu, Junjie Zhang, Xiatian Zhu, Shaogang Gong, and Jingya Wang. Multi-modal multi-platform person re-identification: Benchmark and method. In *ICCV*, pages 10251–10261, 2025. 3, 4

[4] Weizhen He, Yiheng Deng, Shixiang Tang, Qihao Chen, Qingsong Xie, Yizhou Wang, Lei Bai, Feng Zhu, Rui Zhao, Wanli Ouyang, et al. Instruct-reid: A multi-purpose person re-identification task with instructions. In *CVPR*, pages 17521–17531, 2024. 4

[5] He Li, Mang Ye, Ming Zhang, and Bo Du. All in one framework for multimodal re-identification in the wild. In *CVPR*, pages 17459–17469, 2024. 4

[6] Xulin Li, Yan Lu, Bin Liu, Jiaze Li, Qinhong Yang, Tao Gong, Qi Chu, Mang Ye, and Nenghai Yu. Towards anytime retrieval: a benchmark for anytime person re-identification. In *IJCAI*, pages 1467–1475, 2025. 4

[7] Wei Liu, Xin Xu, Hua Chang, Xin Yuan, and Zheng Wang. Mix-modality person re-identification: A new and practical paradigm. *ACM TOMM*, 21(4):1–21, 2025. 4

[8] Huy Nguyen, Kien Nguyen, Akila Pemasiri, Akmal Jahan, Clinton Fookes, and Sridha Sridharan. Ag-vpreid. vir: Bridging aerial and ground platforms for video-based visible-infrared person re-id. *arXiv preprint arXiv:2507.17995*, 2025. 3

[9] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *ECCV*, pages 1–21. Springer, 2022. 2

[10] Aihua Zheng, Zi Wang, Zihan Chen, Chenglong Li, and Jin Tang. Robust multi-modality person re-identification. In *AAAI*, pages 3529–3537, 2021. 3

[11] Wei-Shi Zheng, Junkai Yan, and Yi-Xing Peng. A versatile framework for multi-scene person re-identification. *IEEE TPAMI*, 47(3):1362–1380, 2024. 4

[12] Jialong Zuo, Yongtai Deng, Mengdan Tan, Rui Jin, Dongyue Wu, Nong Sang, Liang Pan, and Changxin Gao. Reid5o: Achieving omni multi-modal person re-identification in a single model. In *NeurIPS*, 2025. 4