

Multi-modal Multi-platform Person Re-Identification: Benchmark and Method

Ruiyang Ha^{1,†} Songyi Jiang^{1,†} Bin Li¹ Bikang Pan¹ Yihang Zhu¹
 Junjie Zhang² Xiatian Zhu³ Shaogang Gong⁴ Jingya Wang^{1,*}
¹ShanghaiTech University, ²The Xi'an Jiaotong-Liverpool University
³University of Surrey ⁴Queen Mary University London

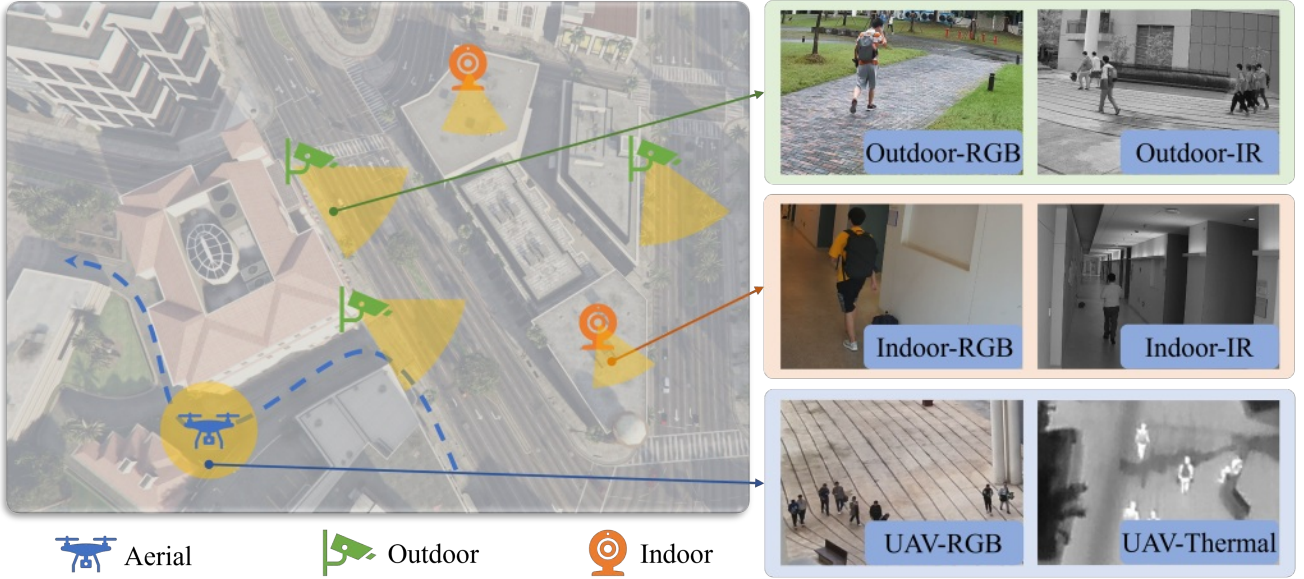


Figure 1. A new MP-ReID dataset is represented by a conceptual diagram, showcasing the inclusion of six ground RGB cameras, six ground infrared cameras, one UAV RGB camera and one UAV thermal camera. The data is collected from a variety of environments, including outdoor and indoor settings, as well as from UAV cameras for aerial perspectives. This integration of diverse data sources and modalities creates a comprehensive and versatile dataset, aimed at enhancing research in multi-modal human perception.

Abstract

Conventional person re-identification (ReID) research is often limited to single-modality sensor data from static cameras, which fails to address the complexities of real-world scenarios where multi-modal signals are increasingly prevalent. For instance, consider an urban ReID system integrating stationary RGB cameras, nighttime infrared sensors, and UAVs equipped with dynamic tracking capabilities. Such systems face significant challenges due to variations in camera perspectives, lighting conditions, and sensor modalities, hindering effective person ReID. To address these challenges, we introduce the MP-ReID benchmark,

a novel dataset designed specifically for multi-modality and multi-platform ReID. This benchmark uniquely compiles data from 1,930 identities across diverse modalities, including RGB, infrared, and thermal imaging, captured by both UAVs and ground-based cameras in indoor and outdoor environments. Building on this benchmark, we introduce Uni-Prompt ReID, a framework with specific-designed prompts, tailored for cross-modality and cross-platform scenarios. Our method consistently outperforms state-of-the-art approaches, establishing a robust foundation for future research in complex and dynamic ReID environments. Our dataset are available at: <https://mp-reid.github.io/>.

[†]Equal contributions.

^{*}Corresponding author.

1. Introduction

Person Re-Identification (ReID)[36, 40, 52] aims to identify and match pedestrians captured by non-overlapping, wide-field cameras. Traditional ReID approaches [2, 4, 19, 24, 34] have primarily relied on static cameras and single-modality data, which limits their ability to handle the dynamic and diverse conditions of real-world environments. However, recent advances in diverse platforms and sensors, including vehicle-mounted cameras[3], unmanned aerial vehicles (UAVs)[30, 46], and wearable cameras[27], offer new opportunities to exploit multi-modality and multi-platform data. These developments enable data collection from various perspectives, environments, and lighting conditions, significantly enriching the data available for ReID systems and enhancing their adaptability in complex scenarios. For instance, consider a 24/7 urban pedestrian ReID system integrating RGB cameras, nighttime infrared sensors, and UAVs with active tracking capabilities. Such a multi-modality, multi-platform framework provides comprehensive coverage, capturing data continuously for uninterrupted monitoring. UAVs, in particular, offer dynamic tracking and flexible viewpoints, while diverse modalities enhance situational awareness. By efficiently fusing data from different platforms and sensors, such systems can achieve a holistic understanding of the monitored environment, improving security and data collection efficiency.

Existing cross-modality datasets[23, 26, 35, 49] are limited to RGB and infrared modalities, collected primarily using static ground cameras, failing to address the complexities of real-world applications. Although UAV-based datasets like AG-ReID[27] and AG-ReID v2[28] have been introduced, they focus only on aerial and ground RGB data and lack modality diversity. The absence of datasets combining multi-modality with multi-platform capabilities highlights a critical gap in the field.

To address this gap, we introduce the Multi-modality Multi-platform ReID (MP-ReID) benchmark, a novel dataset tailored for complex urban environments. As illustrated in Fig. 1, MP-ReID comprises data collected from ground RGB cameras, ground infrared cameras, and UAVs equipped with RGB and thermal cameras, deployed across diverse indoor and outdoor locations. The dataset includes 1,930 identities and 136,156 manually annotated bounding boxes, ensuring robust support for research. Although the dataset size is constrained by the high costs of multi-modality and multi-platform data collection and annotation, MP-ReID surpasses existing cross-modality and cross-platform datasets in scale and diversity (Table 1).

However, simultaneously addressing the challenges posed by modality discrepancies and ground-aerial disparities remains an extremely difficult problem. Existing methods primarily focus on bridging the modality gaps between infrared/RGB or thermal/RGB[6, 9, 15, 16, 29, 31,

33, 37, 39, 41, 42, 45, 49, 50], largely ignoring the additional complexities introduced by diverse capture devices. These complexities include variations in viewpoints, resolutions, and indoor/outdoor environments. Leveraging the advancements in visual foundation models, we incorporate pretrained vision-language models such as CLIP and apply prompt learning to better handle complex scenarios involving multiple modalities and platforms. To this end, we propose Uni-Prompt ReID, a unified prompt learning framework that integrates modality-aware prompts and platform-aware prompts with visual-enhanced context into a unified text prompt structure. By doing so, our model achieves enhanced adaptability to diverse conditions, leading to significant improvements in pedestrian ReID performance across a wide range of challenging environments. In summary, the main contributions of this work are as follows:

1. We introduce MP-ReID, the first multi-modality multi-platform benchmark for person re-identification. MP-ReID integrates data from 14 cameras, including 6 ground-based RGB cameras, 6 ground-based infrared cameras, and a UAV equipped with both RGB and thermal sensors, simulating complex urban environments. This innovative setup enables comprehensive person ReID with unprecedented coverage, significantly improving recognition accuracy across diverse scenarios through the synergistic use of multiple modalities and platforms.
2. To address the inherent challenges of multi-modality and multi-platform ReID tasks, we introduce Uni-Prompt ReID, a unified prompt learning framework. Our approach seamlessly integrates modality-aware and platform-aware prompts with visual-enhanced context into a unified text prompt structure. By leveraging these tailored components, Uni-Prompt ReID effectively bridges the gaps across diverse modalities and platforms, setting a new standard for robustness and adaptability in ReID tasks.
3. We conducted extensive experiments on the newly introduced benchmark, demonstrating the superior performance of our proposed framework. To foster further research and ensure reproducibility, we will publicly release the MP-ReID dataset and provide all code for open access.

2. Related Work

2.1. Person Re-identification Dataset

In the field of ReID, numerous datasets have been developed to address various challenges. VIPeR[11] is the first dataset focus on viewpoint invariant pedestrian recognition. PRID-2011[14], Market-1501[51] and MSMT17[34] expanded on this with more person ID numbers and more camera views. Nevertheless, a single RGB modality strug-

gles to cope with changes in lighting conditions. To address this, the infrared[35] and thermal[26] modalities have been introduced into ReID systems. Recently, several new cross-modal ReID datasets have been proposed. LLCM[49] focuses on the infrared pedestrian images across various illumination scenarios. HITSZ-VCM[23] has launched the first video-based infrared dataset. Additionally, CM-Group[38] has introduced the first cross-modality dataset specifically for group ReID. Lately, advancements of UAV has enabled ReID systems to overcome the limitations of stationary setups. MRP[18] utilizes UAV as a moving platform to introduce the first dataset for UAV ReID system. DRHIT01[12] presents a small-scale dataset with real UAV view. For the subsequent works, PRAI-1581[46] emphasizes a variety of altitudes and P-DESTRE[17] supports long-term ReID. Besides, UAV-Human[21] propose a benchmark for human behavior understanding with UAV. More recently, AG-ReID [27] and AG-ReID v2[28] have introduced a dataset across aerial RGB cameras and ground RGB cameras. Nevertheless, extant research conducted on datasets has exclusively focused on data collected through CCTV or UAVs, and have only considered two modalities. There is a paucity of a dataset that can encompass a wider range of modalities and different collection platforms, which is necessary in complex urban environments and real-world applications.

2.2. Multi-Modality Person Re-Identification

Traditional person ReID primarily focuses on identifying pedestrians within single modality[7, 13, 14, 22, 25, 34, 51, 53], which struggles to overcome the limitations due to varying lighting conditions, typically performing inadequately in low lighting conditions. To address this problem, [26, 35] have introduced infrared and thermal modalities into the ReID system, which led to the emergence of numerous cross-modality ReID studies[6, 33, 39–41, 49, 50]. More specifically, CAJ[39], CAJ₊ [41] and DEEN[49] have made efforts on augmentation strategies to narrow the gap between modalities. AGW[40], MSCLNet[50] and SAAI[6] have attempted to learn features that are invariant across modalities in a latent space. Additionally, to alleviate the laboring efforts, OTLA-ReID[33] proposed an optimal transport based method to assign pseudo labels from visible to infrared modality. However, existing methods focus solely on modality gaps between infrared/RGB or thermal/RGB, overlooking the challenges introduced by diverse capture devices—such as variations in viewpoints, resolutions, and indoor/outdoor environments. A unified framework addressing both cross-modality and cross-platform ReID is urgently needed.

3. Dataset

3.1. Dataset Overview

In urban environments, the inherent challenges—such as complex scenes, architectural occlusions, and high pedestrian mobility—demand the effective integration of diverse sensor modalities and capture platforms. Recognizing that these multi-modality and multi-platform configurations are essential yet underrepresented in existing benchmarks, we introduce MP-ReID, the first dataset tailored to address these practical challenges. Table. 1 provides a quantitative comparison between MP-ReID and several established ReID datasets, underscoring our dataset’s advantages in both scale and diversity of modalities and platforms. We collect images from different platforms and various modalities, making it particularly valuable for the development and evaluation of multi-modality multi-platform person ReID. The MP-ReID consists 1,930 identities and 136,156 images sampled from over 1.2M frames. The images are collected from ground RGB and infrared cameras and UAV RGB and thermal cameras. As illustrated in Fig. 2, our dataset comprises three distinct modalities and three different scenes, with notable disparities between images captured in different modalities and scenes, and it also illustrates the challenges present across various modalities and platforms.

3.2. Comparison with Existing Person ReID Datasets.

Traditional datasets[14, 22, 34, 51, 53] for pedestrian ReID tasks primarily encompass RGB images captured in outdoor environments. The exemplar among these, MSMT17[34], featuring 4,101 unique identities and a total of 126,441 bounding boxes, was acquired through 15 cameras. Cross-modality datasets[23, 26, 35, 49] are devised for more complicated ReID tasks, often incorporating thermal or infrared modalities, and occasionally extending to indoor settings. Existing UAV datasets[17, 21, 27, 46–48] typically exhibit limited scope, with no more than three cameras deployed and restricted to RGB image acquisition for pedestrian ReID purposes. However, our proposed MP-ReID dataset significantly extends beyond conventional limitations. Leveraging data collection from both ground-based cameras and UAV, our dataset contains video recordings capturing RGB, infrared, and thermal modalities across indoor, outdoor, and aerial domains. Noteworthy is the scale of our MP-ReID dataset, featuring 1,930 distinct identities and 136,156 human bounding boxes acquired from 14 cameras. This scale is comparable to that of multi-view traditional datasets, thus establishing our dataset as a substantial resource for ReID research, featuring numerous cameras and human annotations across various platforms and modalities. To ensure robust privacy protection, we applied mosaic techniques to obscure pedestrians’ facial features in

Dataset	IDs.	Cams.	BBox	Scene			Modality			Privacy
				Indoor	Outdoor	Aerial	RGB	Infrared	Thermal	
PRID[14]	200	2	1, 134		✓		✓			
CUHK03[22]	1, 360	2	13, 164		✓		✓			
Market-1501[51]	1, 501	6	32, 668		✓		✓			
MSMT17[34]	4, 101	15	126, 441		✓		✓			
PRAI-1581[46]	1, 581	2	39, 461		✓	✓	✓			
UAV-Human[21]	1, 144	1	41, 290		✓	✓	✓			
P-DESTRE[17]	253	1	> 14.8M		✓	✓	✓			
AG-ReID[27]	388	2	21, 983		✓	✓	✓			✓
AG-ReID v2[28]	1, 615	3	100, 502		✓	✓	✓			✓
G2APS[47]	2, 644	2	260, 559		✓	✓	✓			
RegDB[26]	412	2	8, 240		✓		✓		✓	
SYSU-MM01[35]	491	6	38, 271	✓	✓		✓	✓		
LLCM[49]	1, 064	9	46, 767		✓		✓	✓		✓
HITSZ-VCM[23]	927	12	463, 259	✓	✓		✓	✓		
MP-ReID (Ours)	1, 930	14	136, 156	✓	✓	✓	✓	✓	✓	✓

Table 1. Comparing with Existing Person ReID Datasets. Existing datasets are organized into three distinct categories: (1) conventional RGB-based ReID datasets, (2) UAV ReID datasets, and (3) cross-modality ReID datasets.

all video recordings and permanently deleted the original raw footage to prevent any potential data leakage.

3.3. Data Collection

To maximize dataset diversity and capture a wider range of scenes, we have employed a total of 12 Hikvision DS-2DE3Q140MX-T cameras, recording at a frame rate of 25 fps with a resolution of 1920×1080 for 1 hour. Notably, half of these cameras were configured to operate in full-color mode, while the remaining six were configured for infrared night vision mode. In addition to the ground-based cameras, a DJI Mavic 3T platform, which carried an RGB and a thermal camera placed in parallel with hardware synchronization, to conduct aerial data collection and capture pedestrian activities from an aerial perspective. The resolution of the RGB cameras is 3840×2160 , whereas the resolution of the thermal camera is 640×512 . Both cameras capture simultaneously at a frame rate of 30 fps. Furthermore, the cruising altitudes were set at 5m, 7m, and 10m, capturing videos from varying perspectives and angles ranging from 30 to 80 degrees relative to the ground. The total duration of the collected videos exceeded 13 hours. Our research team has received ethics approval from the Institutional Review Board for this project and the building administration has authorized the recording. All participants have been notified that data collection is solely for research objectives. To enhance privacy protection, we have applied facial blurring to all video recordings and permanently deleted the raw footage.

3.4. Privacy Preserving

Ethical approval for this study was obtained from the Institutional Review Board. We secured all necessary shooting permits for the designated areas and ensured compliance by posting public notices within the shooting zones. All cameras were placed in prominent locations for transparency. Participants entering the shooting zones were informed of the protocol and provided explicit consent. To safeguard data privacy, we meticulously anonymized the released dataset by applying mosaic techniques to obscure pedestrians’ facial features. Furthermore, we permanently deleted the raw footage, retaining only the processed dataset to eliminate any risk of privacy exposure. To ensure responsible use, the dataset will be restricted to non-commercial academic purposes, enforced through signed agreements that explicitly prohibit commercial exploitation.

3.5. Data Annotation

To ensure the acquisition of high-quality pedestrian bounding boxes, we utilized the state-of-the-art YOLOX[10] method to track the aforementioned videos. In total, we obtained over 1,500,000 pedestrian bounding boxes. Furthermore, we employed manual filtering on the obtained bounding boxes, resulting in over 790,000 usable bounding boxes from ground cameras and over 250,000 from UAV’s RGB camera. The process of aligning identities obtained from disparate camera sources necessitated a substantial investment of time and effort. In particular, we found that



Figure 2. Our MP-ReID dataset comprises three distinct modalities and three different scenes, with notable disparities between images captured in different modalities and scenes. We showcase a range of variations to highlight the challenges present in person re-identification. From left to right, selected samples from different scenes and modalities illustrate the disparities between various viewpoints, instances of low resolution, cases of motion blur, and scenarios involving occlusion, respectively. These examples serve to demonstrate the complex nature of the gaps and obstacles within our dataset, emphasizing the diversity and real-world applicability of the MP-ReID benchmark.

due to the low resolution and significant modal differences of the UAV’s thermal camera, the tracking performance of YOLOX[10] in thermal videos was unsatisfactory. To overcome this limitation, for thermal videos, we manually compared each frame with the annotated UAV’s RGB videos and manually delineated bounding boxes. Ultimately, this rigorous procedure yielded 1,930 unique IDs, ensuring the accuracy and reliability of our dataset.

3.6. Data Statistics

In our dataset, each ground RGB camera captures an average of 497 IDs and 7,545 bounding boxes, each ground infrared camera captures an average of 369 IDs and 6,050 bounding boxes, the UAV’s RGB camera captures 341 IDs and 26,046 bounding boxes, and the UAV’s thermal camera captures 474 IDs and 28,543 bounding boxes. There are 346 persons captured by only one camera and 381 persons with less than 10 bounding boxes, most person are captured by 2-8 cameras and have 10-60 bounding boxes. Since outdoor scenes pose significant challenges, such as varying lighting conditions, occlusions, and high pedestrian densities, making them essential for robust ReID performance, it is reasonable to have a larger proportion of outdoor scene data than indoor scenes. The bounding box ratios of persons in outdoor, indoor and aerial view are 51.2%, 8.8% and 40.1%, the bounding box ratios of persons in RGB, infrared and thermal modality are 52.3%, 26.7% and 21.0%. For the test IDs appearing in one modality or platform, we randomly select one image from each camera as the query,

while all data from the other modality or platform are used as the gallery during testing. More statistical details and the data splitting details can be found in the Supplementary Material.

3.7. Limitations

While MP-ReID includes data from multiple modalities and platforms, there are still additional modalities and platforms that could be incorporated, such as the wearable device platforms[28] and the event camera data[1]. We encourage future researchers to supplement our work to some extent.

4. Uni-Prompt ReID

To enhance cross-modality and cross-platform generalization in ReID, we introduce Uni-Prompt ReID, a unified prompt-based learning framework that leverages vision-language alignment in VLMs like CLIP through multi-part text prompt fine-tuning. Building on DAPrompt [8], we incorporate modality and platform information into the text prompt. Additionally, inspired by CoCoOp[54], we enhance learnable prompts by designing a weight-tied meta-network to extract text prompt biases from pedestrian image visual features.

Our framework employs a learnable text prompt structure, as illustrated in Fig. 3, which integrates multi-modality and multi-platform information into prompt learning. The text prompt comprises three distinct components: (1) the Specific Re-ID Prompt, which encodes individual-specific

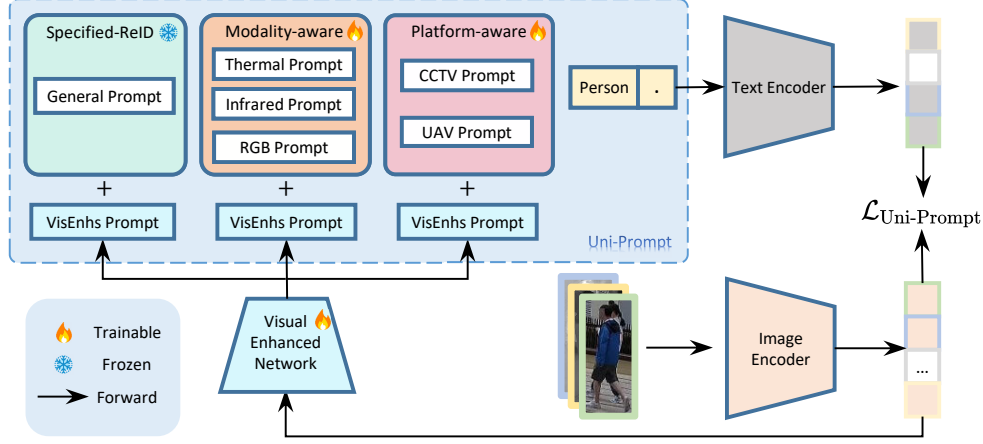


Figure 3. Uni-Prompt ReID divides the learnable context into three parts: Specified-ReID Prompt, Modality-Aware Prompt, and Platform-Aware Prompt. During the training stage, the specified-ReID prompt will be learned first in a warm-up stage and then be frozen with the update of modality-aware prompt and platform-aware prompt. An Inherent Information Embedding Net is also updated with the prompts.

information; (2) the Modality-aware Prompt, which captures modality-specific details; and (3) the Platform-aware Prompt, which incorporates platform-specific context. This design enhances the model’s representational capacity by embedding rich, task-relevant information into the prompt structure.

Moreover, image features contain inherent information that can assist in distinguishing prompts during learning. For example, an infrared image feature may help guide the modality-aware prompt closer to an infrared-specific representation. To use the visual information for enhancing the learnable text prompt, we design a lightweight neural network, named visual-enhanced network, which embeds the image feature into context vectors as a conditioning factor. Let visual-enhanced network $g_\theta(\cdot)$ be parameterized by θ . Given the image feature a , the visual-enhanced context vector is defined as $\sigma := (\sigma_X, \sigma_P, \sigma_M) = g_\theta(a)$. The visual-enhanced text prompt is then defined as $S_m(a) = [S]_m + \sigma_S$ where $[S]_m \in \{X_m, P_m, M_m\}$ represents the m -th text token before enhancement. Here, X, P, M denote specific ReID prompts, modality-aware prompts, and platform-aware prompts, respectively. The visual-enhanced context vector σ is added to the corresponding text prompt to enhance the modality and platform context. During training, the prompt vectors are updated along with the VE-Net parameters θ . Overall, the i -th person prompt is then formulated as:

$$t_i(a) = X_1(a)X_2(a) \dots X_M(a) \\ P_1(a)P_2(a) \dots P_R(a) \\ M_1(a)M_2(a) \dots M_B(a), \text{person}_i. \quad (1)$$

During the text prompt learning phase, we use a two-stage training process. In the first stage, we fix the modality-

aware prompt and platform-aware prompt and use image to text loss and text to image loss proposed in \mathcal{L}_{i2t} and \mathcal{L}_{t2i} from CLIP-ReID [20] to learn the specific Re-ID prompt.

In the second stage, we fix the learned specified Re-ID prompt and use \mathcal{L}_m and \mathcal{L}_p to learn the modality-specific and platform-specific prompt. The modality-specific and platform-specific contrastive losses are separately calculated as:

$$\mathcal{L}_{mi2t}(i) = -\log \frac{\exp(s(V_i, T_{m_i}))}{\sum_{a=1}^B \exp(s(V_a, T_{m_a}))} \quad (2)$$

$$\mathcal{L}_{mt2i}(c_i) = \frac{-1}{|P(c_i)|} \sum_{p \in P(c_i)} \log \frac{\exp(s(V_p, T_{m_{c_i}}))}{\sum_{a=1}^B \exp(s(V_a, T_{m_{c_i}}))} \quad (3)$$

$$\mathcal{L}_{pi2t}(i) = -\log \frac{\exp(s(V_i, T_{p_i}))}{\sum_{a=1}^B \exp(s(V_a, T_{p_a}))} \quad (4)$$

$$\mathcal{L}_{pt2i}(c_i) = \frac{-1}{|P(c_i)|} \sum_{u \in P(c_i)} \log \frac{\exp(s(V_u, T_{p_{c_i}}))}{\sum_{a=1}^B \exp(s(V_a, T_{p_{c_i}}))} \quad (5)$$

where the T_m and T_p are embedded text features with fixed vanilla tokens and specific modality or platform tokens. The number of positive samples is defined as $P(c_i) = \{p \in 1 \dots B : y_p = y_i\}$. The overall loss for Uni-Prompt is expressed as:

$$\mathcal{L}_{\text{Uni-Prompt}} = \mathcal{L}_{mi2t} + \mathcal{L}_{mt2i} + \mathcal{L}_{pi2t} + \mathcal{L}_{pt2i} \quad (6)$$

5. Experiments

5.1. Evaluation Metrics

Following the work of [5, 34, 51], we employed both Cumulated Matching Characteristic (CMC) at Rank-1 and Rank-5. To account for the retrieval of multiple instances and

difficult samples, we use mean average precision (mAP) as the accuracy metrics.

5.2. Implementation Details

We established the benchmark by employing our Uni-Prompt ReID method alongside several representative baseline methods in the field of Visible-Infrared ReID [6, 33, 39–41, 44, 49]. Details of these baselines are provided in the next subsection. To ensure a fair comparison, we conducted experiments on a cluster equipped with eight Nvidia RTX 4090 GPUs each with 24GB of memory. During training, we employed random erasing with a probability of 0.5, random horizontal flipping, and random cropping for data augmentation. All other configurations were set to their default values. For testing, we randomly selected a query image for each identity per camera, and all the images of identities captured by the corresponding camera were used as the gallery. Testing for each model and setting was repeated across 10 trials, and the average metrics were reported. Further details on the training and testing data splitting can be found in Supplementary Materials.

5.3. Experimental Results

To establish a comprehensive benchmark, we compared our Uni-Prompt ReID method with the following baseline approaches. **CAJ** [39] employs a channel-augmented joint learning strategy for color robustness and an enhanced channel-mixed learning approach to address modality variations. **CAJ+** [41] utilizes random channel exchange and a weak-and-strong augmentation joint learning strategy, significantly enhancing its effectiveness. **AGW** [40] incorporates a non-local attention block [43] and generalized-mean pooling [32] with weighted triplet loss for cross-modality visible-infrared tasks. **DEEN** [49] proposes a diverse embedding expansion module with a center-guided pair mining loss for data augmentation in the embedding space to narrow gaps between visible and infrared images. **OTLA-ReID** [33] proposes an Optimal-Transport label assignment module to align infrared and RGB images and employs prediction alignment learning to mitigate the negative effects of inaccurate pseudo-labels. **SAAI** [6] aggregates latent features via pixel-prototype correlation and optimizes inference with a pedestrian affinity module. **CSDN** [44] proposes a CLIP-driven semantic discovery network to integrate text features from both visible and infrared modalities, embedding comprehensive semantics into visual representation.

To validate the various characteristics of our proposed MP-ReID, we designed a total 12 different experimental settings, considering all four image sources: ground RGB, ground infrared, UAV RGB and UAV thermal. It is important to note that, in order to address the complexity and variability of city-scale scenarios, it is essential to fully leverage

the interaction between multiple platforms and the complementarity of different modalities. To this end, we propose four novel cross-modality and cross-platform ReID settings. The results from each experiment were recorded and subsequently categorized into cross-modality, cross-platform, and cross-modality & platform. The mean performance of each method in each experimental category is shown in Table. 2. The findings in the table reveal that most of the cross-modality methods perform well in cross-modality tasks but underperform in the cross modality & platform settings due to the lack of consideration for platform diversity. Our Uni-Prompt ReID method outperforms other state-of-the-art methods, with a 7.87% boost in average Rank-1 accuracy and a 15.42% higher mAP by on our MP-ReID dataset, demonstrating that Uni-Prompt ReID method exhibits strong robustness and high efficiency. Further details of all 12 experimental results are provided in the Supplementary Materials.

5.4. Ablation Study

We conducted ablative analysis on Uni-Prompt ReID using the MP-ReID dataset, categorizing results into cross-modality, cross-platform, and cross-modality & platform settings. Table. 3 confirms its effectiveness in cross-modality & platform scenarios, with detailed results in Supplementary Materials.

Modality-aware Prompt. We conducted a comparative analysis of the Modality-aware prompt and a naive prompt, referred to as the base in Table. 3, which incorporates learnable specific-ReID tokens. Specifically, an average increase of 6.23% in Rank-1 accuracy and a notable 7.09% increase in mAP were observed for cross-modality settings. Additionally, the performance of both cross-platform settings and cross-modality & platform settings was marginally enhanced. These results demonstrate that the Modality-aware prompt significantly improves accuracy in cross-modality tasks without negatively impacting performance in other settings.

Platform-aware Prompt. We enhanced the Modality-aware prompt to a Modality & Platform-aware prompt by adding a learnable platform-aware prompt, improving Rank-1 accuracy and mAP in both cross-platform and cross-modality & platform tasks. Notably, this update surpassed state-of-the-art methods, especially in complex cross-modality & platform scenarios, highlighting the prompt’s effectiveness.

Visual-enhanced Prompt. As demonstrated in Table. 3, the inclusion of the Visual-enhanced Prompt into the Modality & Platform-aware prompt mentioned above leads to further improvement in each experiment and brings a 1.85% Rank-1 accuracy boost in Cross-modality only experiments and a 2.50% increase in those of the cross-modality & platform experiments. What’s more, Uni-

	Cross-platform Only			Cross-modality Only			Cross-modality & platform			Average		
	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP
CAJ [39]	40.36	63.81	29.77	45.34	64.63	26.84	10.62	24.27	7.91	32.11	50.90	21.51
CAJ+ [41]	47.60	70.09	37.53	58.16	76.17	38.52	21.51	41.61	15.79	42.42	62.62	30.61
AGW [40]	53.68	75.88	42.88	51.88	69.32	33.90	19.21	37.94	14.91	41.59	61.05	30.56
DEEN [49]	60.05	79.82	49.02	69.59	85.55	48.88	27.59	51.76	20.07	52.41	72.38	39.33
OTLA-ReID [33]	73.24	85.60	61.85	68.12	84.72	46.57	29.31	53.27	20.66	56.89	74.53	43.03
SAAI [6]	68.11	82.71	53.50	60.42	77.98	41.71	28.40	51.59	20.31	52.31	70.76	38.51
CSDN [44]	31.65	54.16	23.22	41.34	53.90	32.73	12.17	20.49	11.40	28.39	42.85	22.45
Uni-Prompt(Ours)	78.77	88.38	74.29	72.26	85.00	60.81	43.16	60.60	40.26	64.73	77.99	58.45

Table 2. We classified the experimental settings into three categories: cross-modality only, cross-platform only, and both cross-modality and cross-platform. We then compare the average performance of these three categories.

	Cross-platform Only			Cross-modality Only			Cross-modality & platform			Average		
	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP
Base	77.01	81.63	71.19	61.11	70.71	49.68	28.40	38.53	23.08	55.51	63.62	47.98
Base+MP	77.18	84.26	71.34	67.34	81.24	56.77	31.57	46.18	26.91	58.70	70.56	51.67
Base+MP+PP	78.62	88.01	73.68	70.31	81.88	60.19	40.66	56.55	38.57	63.20	75.48	57.48
Base+MP+PP+VE(Full)	78.77	88.38	74.29	72.26	85.00	60.81	43.16	60.60	40.26	64.73	77.99	58.45

Table 3. Ablation study of specifically designed prompts. Here, the Base is the specific ReID prompt learned by the CLIP-ReID loss, and MP, PP, VE represent the modality-aware prompt, platform-aware prompt and visual-enhanced prompt, respectively.

Prompt ReID achieves state-of-the-art performance within the visual-enhanced prompt, highlighting the advantage of this method in the complex scenarios where inconsistencies in visual features across modalities and differences in camera features in cross-platform scenes coexist.

5.5. Result Analysis

As shown in Table. 2, we categorized the total experimental settings into three classes: cross-modality only, cross-platform only, and both cross-modality and cross-platform. A general observation from our results indicates that existing baseline methods struggle to achieve satisfactory performance across these different settings within our dataset. The only scenarios where their performance can be considered acceptable are those that involve either a single cross-modality or a single cross-platform setting, provided that ground camera RGB data is available. This relatively better performance in such cases can be attributed to the smaller disparity introduced when only one type of variation is present. Additionally, the RGB data from ground cameras offers rich visual information, which serves as a crucial contextual aid in mitigating the challenges posed by either cross-modality or cross-platform discrepancies.

However, when it comes to the scenarios that include both cross-modality and cross-platform simultaneously, existing baseline methods exhibit significant shortcomings. This sharp decline is due to the fact that existing ReID datasets do not encompass the highly challenging scenarios as shown in Fig. 2, involving simultaneous cross-modality and cross-platform variations.

This performance gap highlights the necessity of our

proposed MP-ReID benchmark, designed to overcome limitations in current datasets and evaluation settings. Unlike conventional methods, our Uni-Prompt ReID approach unifies multiple prompt mechanisms to handle both cross-modal and cross-platform variations. Specifically, it integrates modality-aware, platform-aware, and visual-enhanced prompts within a well-structured design, effectively capturing and leveraging information across modalities and platforms. This leads to superior performance across multiple tasks, setting new SOTA results and offering a promising solution to real-world multi-modal, multi-platform ReID challenges.

6. Conclusion

In this work, we introduce the MP-ReID dataset, a comprehensive multi-modality, multi-platform ReID benchmark designed to address the limitations of existing datasets in capturing modality and device variations. Compared to previous datasets, MP-ReID offers a broader range of modalities and platforms, making it one of the most extensive benchmarks available. Its diverse variations—including differences in viewpoints, lighting, resolutions, and environments—ensure a more challenging and realistic evaluation for real-world ReID tasks. To tackle these challenges, we propose Uni-Prompt ReID, a unified framework specifically designed for cross-modality and cross-platform ReID. Our approach not only achieves state-of-the-art performance across all experimental setups but also significantly improves the handling of intricacies and domain gaps in multi-modal, multi-platform ReID.

References

- [1] Shafiq Ahmad, Pietro Morerio, and Alessio Del Bue. Person re-identification without identification via event anonymization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11132–11141, 2023. 5
- [2] Xiang Bai, Mingkun Yang, Tengting Huang, Zhiyong Dou, Rui Yu, and Yongchao Xu. Deep-person: Learning discriminative deep features for person re-identification. *Pattern Recognition*, 98:107036, 2020. 2
- [3] Huanjie Chen, Wenjie Cai, Feng Wu, and Qiong Liu. Vehicle-mounted far-infrared pedestrian detection using multi-object tracking. *Infrared Physics & Technology*, 115:103697, 2021. 2
- [4] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 403–412, 2017. 2
- [5] Chanho Eom and Bumsub Ham. Learning disentangled representation for robust person re-identification. *Advances in neural information processing systems*, 32, 2019. 6
- [6] Xingye Fang, Yang Yang, and Ying Fu. Visible-infrared person re-identification via semantic alignment and affinity inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11270–11279, 2023. 2, 3, 7, 8
- [7] Shang Gao, Jingya Wang, Huchuan Lu, and Zimo Liu. Pose-guided visible part matching for occluded person reid. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3
- [8] Chunjiang Ge, Rui Huang, Mixue Xie, Zihang Lai, Shiji Song, Shuang Li, and Gao Huang. Domain adaptation via prompt learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 5
- [9] Yiyuan Ge, Zhihao Chen, Ziyang Wang, Jiaju Kang, and Mingya Zhang. Spectral enhancement and pseudo-anchor guidance for infrared-visible person re-identification. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025. 2
- [10] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. YoloX: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 4, 5
- [11] Douglas Gray, Shane Brennan, and Hai Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *Proc. IEEE international workshop on performance evaluation for tracking and surveillance (PETS)*, pages 1–7, 2007. 2
- [12] Aleksei Grigorev, Zhihong Tian, Seungmin Rho, Jianxin Xiong, Shaohui Liu, and Feng Jiang. Deep person re-identification in uav images. *EURASIP Journal on Advances in Signal Processing*, 2019:1–10, 2019. 3
- [13] Lingxiao He, Xingyu Liao, Wu Liu, Xinchun Liu, Peng Cheng, and Tao Mei. Fastreid: A pytorch toolbox for general instance re-identification. *arXiv preprint arXiv:2006.02631*, 2020. 3
- [14] Martin Hirzer, Csaba Belezai, Peter M Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In *Image Analysis: 17th Scandinavian Conference, SCIA 2011, Ystad, Sweden, May 2011. Proceedings 17*, pages 91–102. Springer, 2011. 2, 3, 4
- [15] Yan Jiang, Hao Yu, Xu Cheng, Haoyu Chen, Zhaodong Sun, and Guoying Zhao. From laboratory to real world: A new benchmark towards privacy-preserved visible-infrared person re-identification. *arXiv preprint arXiv:2503.12232*, 2025. 2
- [16] Asaf Joseph and Shmuel Peleg. Clothes-changing person re-identification based on skeleton dynamics. *arXiv preprint arXiv:2503.10759*, 2025. 2
- [17] SV Aruna Kumar, Ehsan Yaghoubi, Abhijit Das, BS Harish, and Hugo Proença. The p-destre: A fully annotated dataset for pedestrian detection, tracking, and short/long-term re-identification from aerial devices. *IEEE Transactions on Information Forensics and Security*, 16:1696–1708, 2020. 3, 4
- [18] Ryan Layne, Timothy M Hospedales, and Shaogang Gong. Investigating open-world person re-identification using a drone. pages 225–240, 2015. 3
- [19] Dangwei Li, Xiaotang Chen, Zhang Zhang, and Kaiqi Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 384–393, 2017. 2
- [20] Siyuan Li, Li Sun, and Qingli Li. Clip-reid: exploiting vision-language model for image re-identification without concrete text labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1405–1413, 2023. 6
- [21] Tianjiao Li, Jun Liu, Wei Zhang, Yun Ni, Wenqian Wang, and Zhiheng Li. Uav-human: A large benchmark for human behavior understanding with unmanned aerial vehicles. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16266–16275, 2021. 3, 4
- [22] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deep-reid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014. 3, 4
- [23] Xinyu Lin, Jinxing Li, Zeyu Ma, Huafeng Li, Shuang Li, Kaixiong Xu, Guangming Lu, and David 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. 2, 3, 4
- [24] Zimo Liu, Jingya Wang, Shaogang Gong, Huchuan Lu, and Dacheng Tao. Deep reinforcement active learning for human-in-the-loop person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6122–6131, 2019. 2
- [25] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 3
- [26] Dat Tien Nguyen, Hyung Gil Hong, Ki Wan Kim, and Kang Ryoung Park. Person recognition system based on a

- combination of body images from visible light and thermal cameras. *Sensors*, 17(3):605, 2017. 2, 3, 4
- [27] Huy Nguyen, Kien Nguyen, Sridha Sridharan, and Clinton Fookes. Aerial-ground person re-id. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 2585–2590. IEEE, 2023. 2, 3, 4
- [28] Huy Nguyen, Kien Nguyen, Sridha Sridharan, and Clinton Fookes. Ag-reid. v2: Bridging aerial and ground views for person re-identification. *IEEE Transactions on Information Forensics and Security*, 2024. 2, 3, 4, 5
- [29] Ke Niu, Haiyang Yu, Mengyang Zhao, Teng Fu, Siyang Yi, Wei Lu, Bin Li, Xuelin Qian, and Xiangyang Xue. Chatreid: Open-ended interactive person retrieval via hierarchical progressive tuning for vision language models. *arXiv preprint arXiv:2502.19958*, 2025. 2
- [30] Daniel Organisciak, Brian KS Isaac-Medina, Matthew Poyser, Shanfeng Hu, Toby P Breckon, and Hubert PH Shum. Uav-reid: A benchmark on unmanned aerial vehicle re-identification. *CoRR*, 2021. 2
- [31] Zhiqi Pang, Lingling Zhao, Yang Liu, Gaurav Sharma, and Chunyu Wang. Inter-modality similarity learning for unsupervised multi-modality person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 2
- [32] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning cnn image retrieval with no human annotation. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1655–1668, 2018. 7
- [33] Jiangming Wang, Zhizhong Zhang, Mingang Chen, Yi Zhang, Cong Wang, Bin Sheng, Yanyun Qu, and Yuan Xie. Optimal transport for label-efficient visible-infrared person re-identification. In *European Conference on Computer Vision*, pages 93–109. Springer, 2022. 2, 3, 7, 8
- [34] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 79–88, 2018. 2, 3, 4, 6
- [35] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. Rgb-infrared cross-modality person re-identification. In *Proceedings of the IEEE international conference on computer vision*, pages 5380–5389, 2017. 2, 3, 4
- [36] Di Wu, Si-Jia Zheng, Xiao-Ping Zhang, Chang-An Yuan, Fei Cheng, Yang Zhao, Yong-Jun Lin, Zhong-Qiu Zhao, Yong-Li Jiang, and De-Shuang Huang. Deep learning-based methods for person re-identification: A comprehensive review. *Neurocomputing*, 337:354–371, 2019. 2
- [37] Ruiqi Wu, Bingliang Jiao, Wenxuan Wang, Meng Liu, and Peng Wang. Enhancing visible-infrared person re-identification with modality-and instance-aware visual prompt learning. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pages 579–588, 2024. 2
- [38] Jianghao Xiong and Jianhuang Lai. Similarity metric learning for rgb-infrared group re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13662–13671, 2023. 3
- [39] Mang Ye, Weijian Ruan, Bo Du, and Mike Zheng Shou. Channel augmented joint learning for visible-infrared recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13567–13576, 2021. 2, 3, 7, 8
- [40] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):2872–2893, 2021. 2, 3, 7, 8
- [41] Mang Ye, Zesen Wu, Cuiqun Chen, and Bo Du. Channel augmentation for visible-infrared re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2, 3, 7, 8
- [42] Qingze Yin, Guodong Ding, Tongpo Zhang, and Yumei Gong. Robust fine-grained learning for cloth-changing person re-identification. *Mathematics*, 13(3):429, 2025. 2
- [43] Hong-Xing Yu, Ancong Wu, and Wei-Shi Zheng. Cross-view asymmetric metric learning for unsupervised person re-identification. In *Proceedings of the IEEE international conference on computer vision*, pages 994–1002, 2017. 7
- [44] Xiaoyan Yu, Neng Dong, Liehuang Zhu, Hao Peng, and Dapeng Tao. Clip-driven semantic discovery network for visible-infrared person re-identification. *IEEE Transactions on Multimedia*, 2025. 7, 8
- [45] Chao Yuan, Guiwei Zhang, Changxiao Ma, Tianyi Zhang, and Guanglin Niu. From poses to identity: Training-free person re-identification via feature centralization. *arXiv preprint arXiv:2503.00938*, 2025. 2
- [46] Shizhou Zhang, Qi Zhang, Yifei Yang, Xing Wei, Peng Wang, Bingliang Jiao, and Yanning Zhang. Person re-identification in aerial imagery. *IEEE Transactions on Multimedia*, 23:281–291, 2020. 2, 3, 4
- [47] Shizhou Zhang, Qingchun Yang, De Cheng, Yinghui Xing, Guoqiang Liang, Peng Wang, and Yanning Zhang. Ground-to-aerial person search: Benchmark dataset and approach. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 789–799, 2023. 4
- [48] Shizhou Zhang, Wenlong Luo, De Cheng, Qingchun Yang, Lingyan Ran, Yinghui Xing, and Yanning Zhang. Cross-platform video person reid: A new benchmark dataset and adaptation approach. In *European Conference on Computer Vision*, pages 270–287. Springer, 2024. 3
- [49] Yukang Zhang and Hanzi Wang. Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2153–2162, 2023. 2, 3, 4, 7, 8
- [50] Yiyuan Zhang, Sanyuan Zhao, Yuhao Kang, and Jianbing Shen. Modality synergy complement learning with cascaded aggregation for visible-infrared person re-identification. In *European Conference on Computer Vision*, pages 462–479. Springer, 2022. 2, 3
- [51] Liang Zheng, Liye Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015. 2, 3, 4, 6

- [52] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016. [2](#)
- [53] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE international conference on computer vision*, pages 3754–3762, 2017. [3](#)
- [54] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825, 2022. [5](#)

Multi-modal Multi-platform Person Re-Identification: Benchmark and Method

Supplementary Material

A. Distractors

In the real-world scene, some individuals may only appear in one camera view, and we designate these as distractors. Table 5 illustrates the performance of each baseline method after incorporating distractors, where the number of distractors is set to 10% of the gallery. Here, we only aim to show the impact of distractors on the results, which is not the primary focus of our work. Therefore, we have selected experimental results from six representative settings. Naturally, the presence of distractors inevitably impacts the performance of models, leading to varying degrees of degradation across different settings for each baseline method.

B. Details of MP-ReID

Here, we show some details of the MP-ReID datasets, including the training and testing data splitting and some statistics analysis.

B.1. Data Splitting

We designed six sets of experiments for ground RGB, ground infrared, UAV RGB, UAV thermal data. For each set, we shuffle pedestrians simultaneously captured by two modal sensors or platforms and randomly select about 2/3 of the IDs as the training set. The remaining part is used as the test set, and the pedestrians captured by only one modal sensor or platform are used as distractors. One exception is that, due to the large volume of ground RGB and ground infrared data and the expensive data annotation in real scenarios, we only randomly select 1/3 as the training set in $G_R \leftrightarrow G_I$. For the test IDs appearing in one modality or platform, we randomly select one image from each camera as the query, while all data from the other modality or platform are used as the gallery during testing.

B.2. Data Statistics

In our dataset, each ground RGB camera captures an average of 497 IDs and 7,545 bounding boxes, each ground infrared camera captures an average of 369 IDs and 6,050 bounding boxes, the UAV’s RGB camera captures 341 IDs and 26,046 bounding boxes, and the UAV’s thermal camera captures 474 IDs and 28,543 bounding boxes. There are 346 persons captured by only one camera and 381 persons with less than 10 bounding boxes, most person are captured by 2-8 cameras and have 10-60 bounding boxes. Since outdoor scenes pose significant challenges, such as varying lighting conditions, occlusions, and high pedestrian densities, making them essential for robust ReID performance, it is reasonable to have a larger proportion of outdoor scene data

than indoor scenes.

As shown in Fig. 5a, Fig. 5b, the bounding box ratios of persons in outdoor, indoor and aerial view are 51.2%, 8.8% and 40.1%, the bounding box ratios of persons in RGB, infrared and thermal modality are 52.3%, 26.7% and 21.0%.

Furthermore, we counted the number of cameras that captured the same pedestrian and the bounding box of each pedestrian. As shown in Fig. 6a and Fig. 6b, there are over 870 persons captured by at least 3 cameras and only 346 persons appear once, which are considered as distractors in the later experiments. And the distribution of person IDs shows that most IDs in the dataset have over 25 bounding boxes, which is very enough to the ReID task.

C. Experiment

Furthermore, we categorized these 12 experimental settings into three classes: cross-modal only, cross-platform only, and both cross-modal and cross-platform. Specifically, cross-platform only: $U_R \rightarrow G_R$, $G_R \rightarrow U_R$; cross-modal only: $G_I \rightarrow G_R$, $G_R \rightarrow G_I$, $U_T \rightarrow U_R$, $U_R \rightarrow U_T$; cross-modal & platform: $U_R \rightarrow G_I$, $G_I \rightarrow U_R$, $U_T \rightarrow G_R$, $G_R \rightarrow U_T$, $U_T \rightarrow G_I$, $G_I \rightarrow U_T$. We show the result of total 12 experimental settings in Table. 6 in detail. What’s more, the performance in each setting for the ablation study can be found in Table. 7.

D. License

The MP-ReID Dataset will be freely available, under the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 (CC BY-NC-ND 4.0) International license.

E. Visualization of Retrieval Results

In Fig. 7, 8 and 9 we visually show the top 10 retrieval results for the tasks of $G_I \rightarrow U_R$ (cross-modal & cross-platform), $U_R \rightarrow G_R$ (cross-platform only), and $U_R \rightarrow U_T$ (cross-modal only) respectively. We designate OTLA-ReID[29] as the representative baseline method due to its superior overall performance. The green rectangles indicate correctly retrieved results, the red ones indicate wrongly retrieved results and the blue ones represent the distractors. As shown, we can observe that our proposed MP-ReID poses a significant challenge to the existing algorithm, particularly after the inclusion of distractors. For instance, as depicted in Fig. 7, we can see that OTLA-ReID exhibits high error rates, which is also heavily influenced by distractors. This arises from the inherent challenges of handling a cross-modal task compounded by various introduced by im-



Figure 4. We provide 6 examples showing images of the same IDs in different scenes and various modalities within our MP-ReID. From left to right, indoor RGB, outdoor RGB, UAV RGB, indoor infrared, outdoor infrared and UAV thermal are shown for each ID, respectively.

Table 4. **Dataset splitting of MP-ReID.** U_T , U_R , G_I and G_R stand for UAV thermal, UAV RGB, ground infrared and ground RGB, respectively. In the **Test** part, the number on the left of '/' corresponds to ' \rightarrow ' direction in the **Setting** part and the right corresponds to ' \leftarrow '. And $BBOX_q$, $BBOX_g$ exhibits the division of query and gallery bounding boxes.

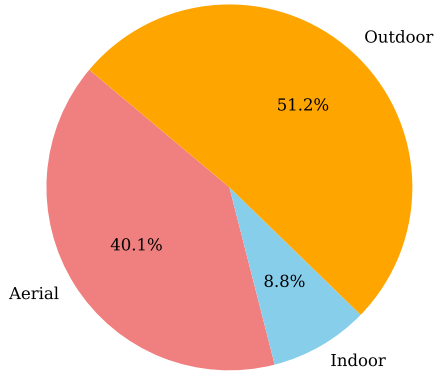
Setting		$U_T \leftrightarrow G_R$	$U_T \leftrightarrow U_R$	$U_T \leftrightarrow G_I$	$G_I \leftrightarrow U_R$	$G_I \leftrightarrow G_R$	$G_R \leftrightarrow U_R$
Train	ID	312	226	274	197	501	227
	BBox	29,670	35,563	24,308	20,045	26,265	24,978
Test	ID	155	112	136	98	1,001	113
	$BBOX_q$	155/329	112/112	136/235	177/98	1,481/1,689	239/113
	$BBOX_g$	4,940/8,373	7,266/7,056	3,777/7,957	8,613/3,109	26,815/23,852	8,600/3,273
	$BBOX_{Distractor}$	30427/402	14/4690	24826/3973	2712/27865	4393/283	17/34447

ages originating from different platforms. From Fig. 8, it becomes evident that the unique perspective provided by the UAV present considerable obstacles for the existing baseline method, which are not explicitly designed to accommodate such intricacies. In Fig. 9, we present the retrieval results from two different modalities obtained by the UAV. Due to the inherent characteristics of the UAV, such as low resolution imaging and mobility, aligning between the RGB and thermal modalities poses additional difficulties. In summary, the aforementioned points highlight the necessity of our proposed MP-ReID, as it can provides strong support for the development of more robust algorithm capable of handling broader ranges of scenes and modalities.

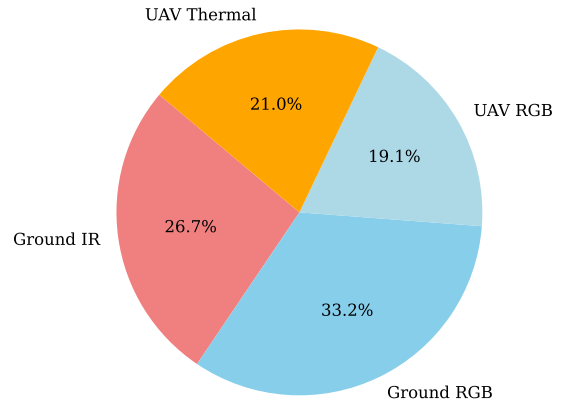
F. Social Impact

The development and deployment of multi-modality multi-platform person ReID dataset carry significant social im-

pact. Incorporating data from various modalities and platforms can promote the development of more accurate and robust personal identification technologies. It greatly aids in crowd analysis, urban planning, and traffic management, thereby advancing the development of smart cities. However, the dataset carries the risk of being targeted for attacks, thereby raising concerns about privacy breaches.

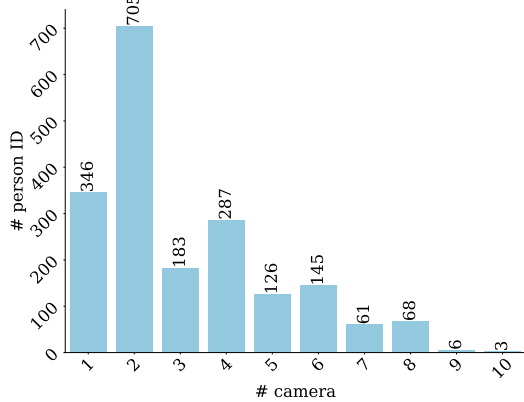


(a) The proportion of bounding boxes in different scenes.

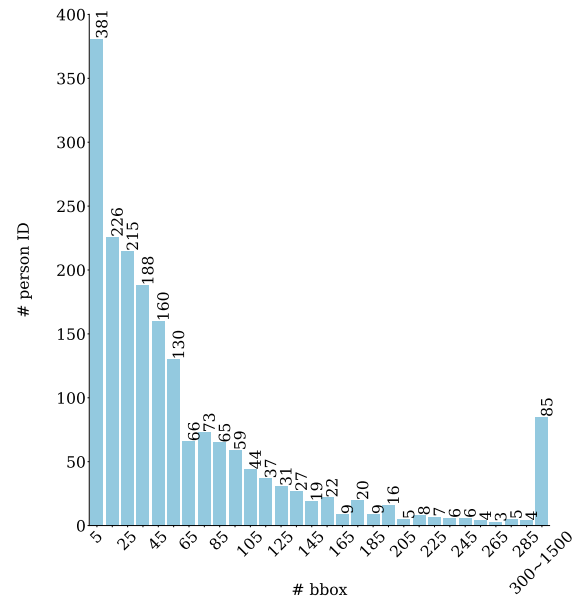


(b) The proportion of bounding boxes in different modalities.

Figure 5. **Bounding box analysis of the MP-ReID dataset.** (a) shows the image proportion in different scenes, and (b) shows the image proposition in different modalities.



(a) Distribution of person IDs Captured by different camera numbers.



(b) Distribution of person IDs across the number of bounding boxes.

Figure 6. Statistical analysis of the MP-ReID dataset.

Method	$U_T \rightarrow G_R$			$U_T \rightarrow G_I$			$U_R \rightarrow U_T$		
	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP
CAJ	2.26	10.32	2.58	1.69	7.43	1.93	22.14	45.80	13.25
CAJ ₊	9.68	26.71	7.20	10.00	25.51	8.41	32.77	61.07	21.34
AGW	11.03	25.55	8.11	11.10	23.09	8.34	24.20	48.84	15.08
DEEN	17.74	40.84	13.59	16.62	32.87	11.00	56.61	79.64	37.76
OTLA-ReID	21.29	41.29	12.89	15.44	36.76	11.17	54.46	75.00	35.72
SAAI	18.93	38.87	12.76	19.94	41.27	14.04	29.04	62.07	21.49
CSDN	10.42	34.94	11.66	7.61	10.01	6.75	15.98	25.30	8.97

Method	$G_I \rightarrow U_R$			$U_R \rightarrow G_R$			$G_I \rightarrow G_R$		
	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP
CAJ	23.50	44.86	16.88	35.31	56.99	23.97	67.43	82.25	37.36
CAJ ₊	41.24	61.98	28.86	40.62	62.48	32.76	83.07	91.23	53.19
AGW	33.11	56.16	25.95	47.17	69.12	34.17	78.80	90.23	51.24
DEEN	44.75	68.53	30.74	54.60	75.22	44.14	84.56	92.57	57.40
OTLA-ReID	39.55	70.06	31.89	69.91	85.84	44.14	82.65	92.44	56.32
SAAI	43.68	66.03	32.42	65.88	77.56	51.86	84.82	92.31	58.59
CSDN	6.34	17.22	14.75	27.54	50.12	19.66	68.72	81.38	37.64

Table 5. The results of distractors added dataset for all baseline methods. Both rank accuracy (%) and mAP(%) are reported.

Query

Top-10 Rank List



Figure 7. Visualization of OTLA-ReID retrieval results. Query images are from ground infrared cameras and gallery images are from UAV RGB cameras. Green, red and blue rectangles indicate correct, wrong retrieval results and distractors, respectively.

Method	$U_R \rightarrow G_R$			$G_R \rightarrow U_R$			$G_I \rightarrow G_R$			$G_R \rightarrow G_I$		
	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP
CAJ	39.12	63.14	30.40	41.59	64.48	29.14	68.68	82.75	38.19	68.32	81.03	40.80
CAJ ₊	41.47	66.57	36.14	53.72	73.60	38.92	84.60	91.57	54.25	77.05	87.38	53.16
AGW	48.53	71.96	41.12	58.83	79.79	44.63	80.38	90.75	52.42	75.60	86.90	51.92
DEEN	56.96	77.45	49.95	63.14	82.18	48.08	85.89	92.82	57.57	79.05	88.47	56.59
OTLA-ReID	74.51	83.33	64.55	71.97	87.87	59.15	84.13	92.57	57.49	78.51	89.17	55.82
SAAI	68.87	81.24	53.88	67.34	84.17	53.11	86.06	92.73	59.89	80.42	89.99	57.84
CSDN	29.41	53.92	23.21	33.89	54.39	23.23	71.34	84.92	42.86	76.17	87.81	56.80
Ours	77.20	87.62	74.16	80.34	89.13	74.41	86.25	92.72	71.21	81.77	89.45	69.30

	$U_T \rightarrow U_R$			$U_R \rightarrow U_T$			$U_R \rightarrow G_I$			$G_I \rightarrow U_R$		
	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP
CAJ	20.62	46.25	14.30	23.75	48.75	14.08	28.88	57.08	19.65	24.92	46.67	17.57
CAJ ₊	36.61	62.32	24.56	34.38	63.39	22.10	43.60	71.01	32.48	43.67	64.97	30.42
AGW	25.18	48.30	15.25	26.34	51.34	16.00	35.73	64.16	28.39	34.80	57.47	27.42
DEEN	55.36	78.57	40.55	58.04	82.32	40.79	46.72	71.36	32.20	46.85	75.06	37.64
OTLA-ReID	53.57	81.25	35.87	56.25	75.89	37.08	56.18	76.40	39.86	41.81	72.88	33.34
SAAI	37.16	63.52	24.88	38.03	65.66	24.21	47.74	72.31	32.53	43.68	68.06	32.95
CSDN	10.71	24.11	15.43	7.14	18.76	16.14	24.72	44.45	13.44	16.95	28.25	10.10
Ours	56.32	80.33	52.46	58.70	77.51	50.26	61.28	79.24	45.88	61.67	78.12	49.77

	$U_T \rightarrow G_R$			$G_R \rightarrow U_T$			$U_T \rightarrow G_I$			$G_I \rightarrow U_T$		
	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP
CAJ	2.58	11.81	2.84	3.13	11.91	2.96	2.13	8.46	1.87	2.09	9.66	2.55
CAJ ₊	10.71	29.35	7.78	8.42	25.93	6.39	12.06	28.97	9.12	10.6	29.45	8.56
AGW	11.74	27.61	8.61	8.97	26.11	7.83	13.16	25.66	9.11	10.85	26.64	8.29
DEEN	20.19	45.61	14.12	18.92	43.4	14.25	17.65	36.47	11.78	15.23	38.64	10.40
OTLA-ReID	23.87	42.58	13.62	17.33	41.64	12.29	18.38	39.71	11.75	18.30	46.38	13.08
SAAI	20.39	42.41	13.47	19.55	42.33	15.31	20.25	43.30	14.55	18.77	41.14	13.04
CSDN	11.58	16.10	10.62	8.82	12.48	8.47	8.94	11.15	7.39	7.98	10.53	6.38
Ours	34.71	50.48	34.14	31.82	48.16	37.23	35.39	50.21	34.65	34.08	57.39	39.89

Table 6. The results of all baseline methods. Both rank accuracy (%) and mAP(%) are reported. U_T , U_R , G_I and G_R stand for UAV thermal, UAV RGB, ground infrared and ground RGB, respectively.

Method	$U_R \rightarrow G_R$			$G_R \rightarrow U_R$			$G_I \rightarrow G_R$			$G_R \rightarrow G_I$		
	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP
Base	74.52	80.79	71.04	79.50	82.46	71.33	85.55	90.55	69.01	82.88	88.67	67.20
Base+MS Prompt	74.37	80.71	71.36	79.98	87.81	71.31	85.62	91.16	70.51	83.94	88.96	70.59
Base+PM Prompt	77.26	87.20	73.34	79.97	88.81	74.21	86.03	91.66	70.78	84.70	89.06	70.51
Base+IE	74.77	81.21	71.22	79.87	86.74	71.31	85.62	90.77	70.38	83.03	89.01	69.21
Base+MS Prompt+IE	74.81	84.21	72.95	80.10	88.66	72.67	85.88	92.19	71.29	84.12	89.41	71.24
Base+PM Prompt+IE	77.20	87.62	74.16	80.34	89.13	74.41	86.25	92.72	71.24	87.77	89.45	69.30

	$U_T \rightarrow U_R$			$U_R \rightarrow U_T$			$U_R \rightarrow G_I$			$G_I \rightarrow U_R$		
	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP
Base	34.11	45.34	31.40	41.88	58.27	31.11	43.40	54.44	35.41	50.21	58.48	41.87
Base+MS Prompt	43.71	72.23	42.91	56.10	72.62	43.07	50.51	63.91	40.62	51.94	65.90	43.51
Base+PM Prompt	54.76	74.31	51.49	55.76	72.49	47.98	59.52	72.64	41.59	59.65	74.21	49.68
Base+IE	40.19	66.35	38.28	47.17	63.69	42.73	48.57	61.97	40.26	51.46	63.89	42.67
Base+MS Prompt+IE	47.67	74.50	44.63	56.96	74.80	47.48	52.02	67.65	45.91	54.61	70.33	44.89
Base+PM Prompt+IE	56.32	80.33	52.46	58.70	77.51	50.26	61.28	79.24	45.88	61.67	78.12	49.77

	$U_T \rightarrow G_R$			$G_R \rightarrow U_T$			$U_T \rightarrow G_I$			$G_I \rightarrow U_T$		
	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP
Base	21.04	28.53	17.81	19.61	30.02	14.87	17.14	28.09	14.51	19.02	31.64	14.03
Base+MS Prompt	23.67	37.11	22.51	19.87	34.39	18.84	22.49	32.52	16.67	20.97	43.23	19.34
Base+PM Prompt	32.86	45.09	30.87	30.23	44.84	36.29	31.25	47.93	34.46	30.45	54.59	38.51
Base+IE	23.16	36.57	20.23	19.89	34.01	16.67	17.52	30.08	16.64	20.49	35.57	19.28
Base+MS Prompt+IE	24.89	41.47	30.02	29.21	41.76	26.24	27.13	37.24	28.88	24.41	46.58	28.80
Base+PM Prompt+IE	34.71	50.48	34.14	31.82	48.16	37.23	35.39	50.21	34.65	34.08	57.39	39.89

Table 7. The results of all Ablation studys. Both rank accuracy (%) and mAP(%) are reported. U_T , U_R , G_I and G_R stand for UAV thermal, UAV RGB, ground infrared and ground RGB, respectively.

Query

Top-10 Rank List



Figure 8. Visualization of OTLA-ReID retrieval results. Query images are from UAV RGB cameras and gallery images are from ground RGB cameras. Green, red and blue rectangles indicate correct, wrong retrieval results and distractors, respectively.

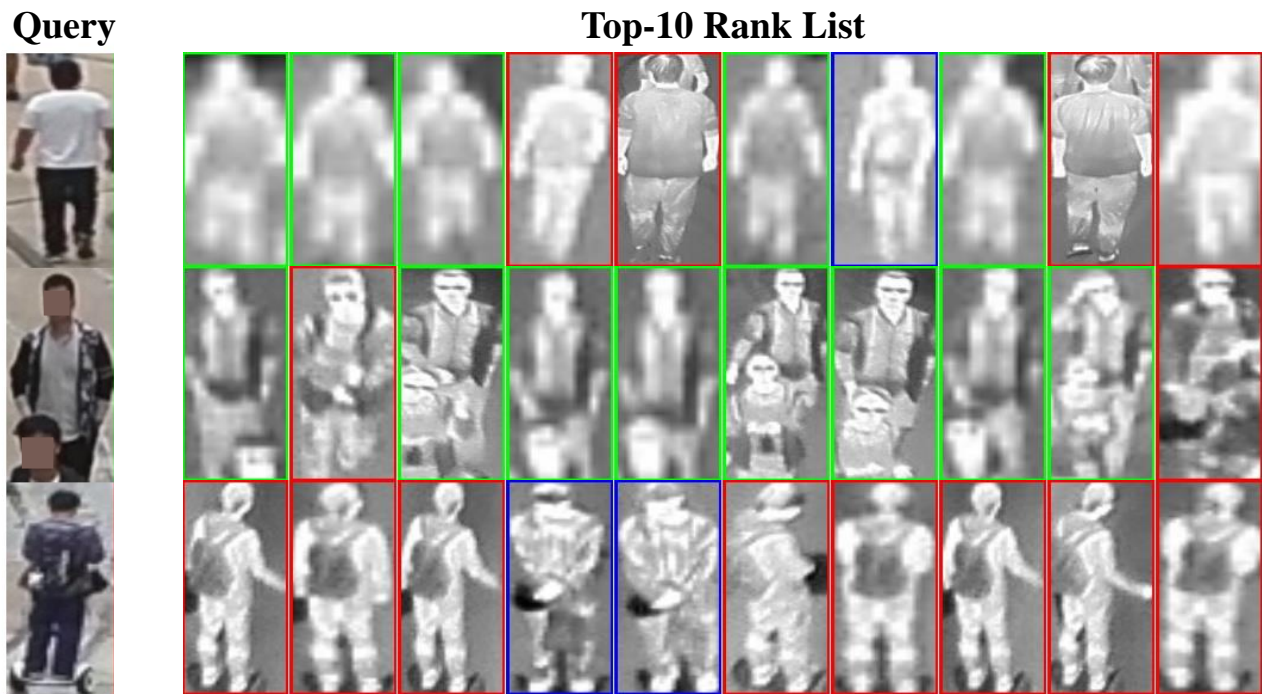


Figure 9. Visualization of OTLA-ReID retrieval results. Query images are from UAV RGB cameras and gallery images are from UAV thermal cameras. Green, red and blue rectangles indicate correct, wrong retrieval results and distractors, respectively.