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

E. Visualization of Retrieval Results

In Fig. 7, 8 and 9 we visually show the top 10 retreival 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 handing a cross-modal task compunded 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.

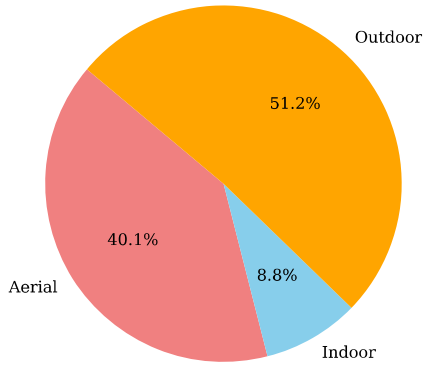
Table 4. **Dataset spliting of MP-ReID.** U_T, U_R, G_I and G_R stand for UAV thermal, UAV RGB, ground infrared and ground RGB, respectively. In the **Test** part, the number on the left of '/' corresponds to '$\rightarrow$' direction in the **Setting** part and the right corresponds to '$\leftarrow$'. 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
	$\text{BBox}_\text{Distractor}$	30427/402	14/4690	24826/3973	2712/27865	4393/283	17/34447

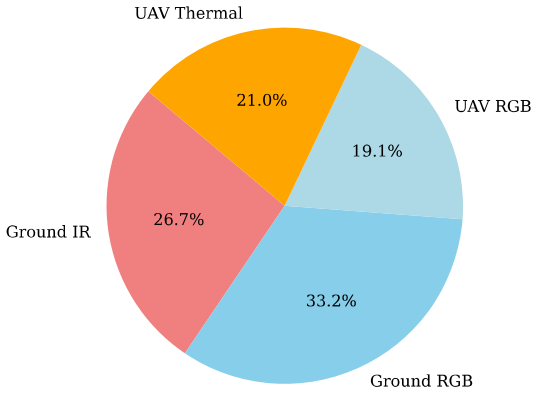
ages originating from different platforms. From Fig.8, it becomes evident that the unique perspective provided by the UAV present considerable obstacles for the existing baseline method , which are not explictly designed to accommodate such intricacies. In Fig. 9, we present the retrieval results from 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 impact. 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.

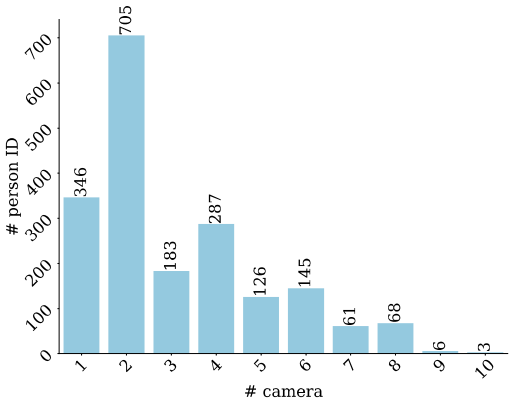


(a) The proportion of bounding boxes in different scenes.

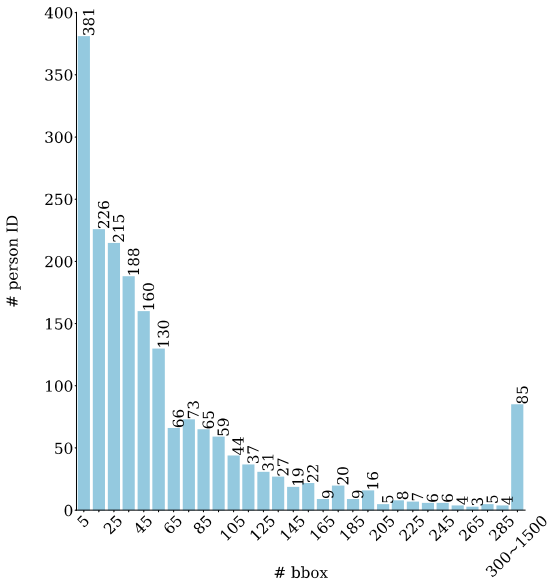


(b) The proportion of bounding boxes in different modalities.

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.

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.

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.

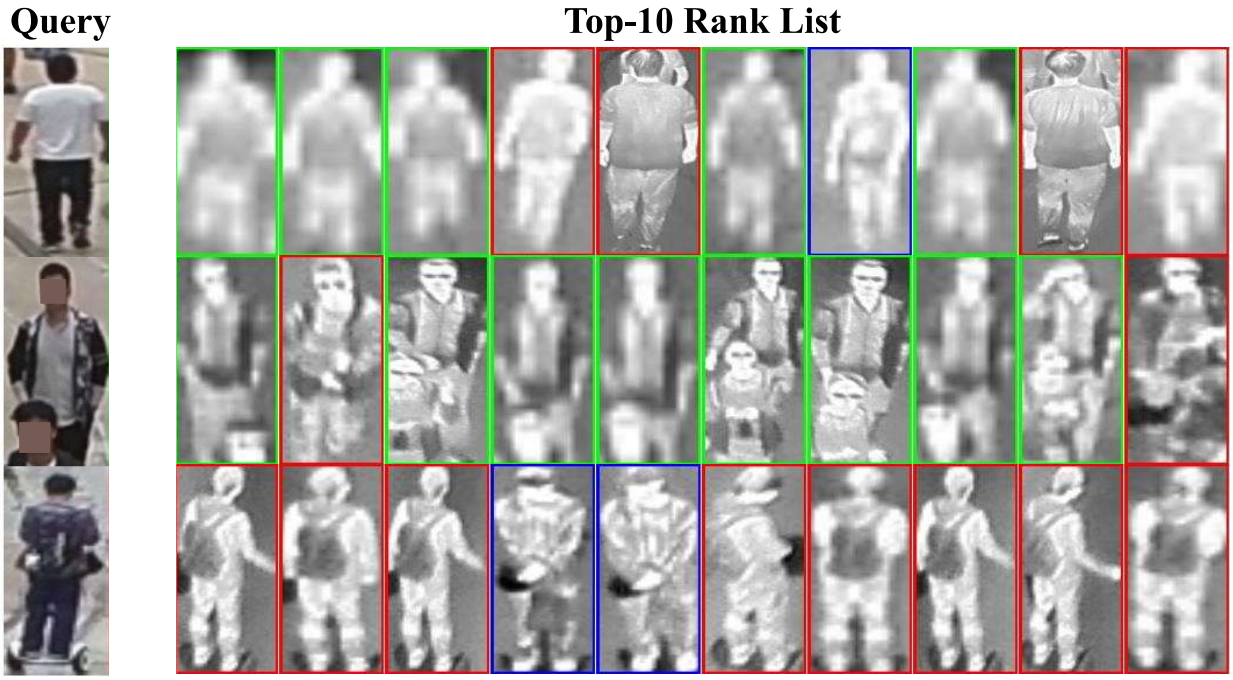


Figure 9. Visualization of OTLA-ReID retrieval results. Query images are from UAV RGB cameras and gallery images are from UAV thermal cameras. Green, red and blue rectangles indicate correct, wrong retrieval results and distractors, respectively.