Multi-Memory Matching for Unsupervised Visible-Infrared Person Re-Identification

Jiangming Shi[1], Xiangbo Yin[2], Yeyun Chen[1], Yachao Zhang[3], Zhizhong Zhang[4,5], Yuan Xie[4⋆], and Yanyun Qu[1,2⋆]

[1] Institute of Artificial Intelligence, Xiamen University, Xiamen, China
[2] School of Informatics, Xiamen University, Xiamen, China
[3] Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
[4] School of Computer Science and Technology, East China Normal University, Shanghai, China
[5] Shanghai Key Laboratory of Computer Software Evaluating and Testing, Shanghai, China
jiangming.shi@outlook.com, yyqu@xmu.edu.cn
{yxie,zzzhang}@cs.ecnu.edu.cn
https://github.com/shijiangming1/MMM

Abstract. Unsupervised visible-infrared person re-identification (USL-VI-ReID) is a promising yet highly challenging retrieval task. The key challenges in USL-VI-ReID are to accurately generate pseudo-labels and establish pseudo-label correspondences across modalities without relying on any prior annotations. Recently, clustered pseudo-label methods have gained more attention in USL-VI-ReID. However, most existing methods don't fully exploit the intra-class nuances, as they simply utilize a single memory that represents an identity to establish cross-modality correspondences, resulting in noisy cross-modality correspondences. To address the problem, we propose a Multi-Memory Matching (MMM) framework for USL-VI-ReID. We first design a simple yet effective Cross-Modality Clustering (CMC) module to generate the pseudo-labels through clustering together both two modality samples. To associate cross-modality clustered pseudo-labels, we design a Multi-Memory Learning and Matching (MMLM) module, ensuring that optimization explicitly focuses on the nuances of individual perspectives and establishes reliable cross-modality correspondences. Finally, we design a Soft Cluster-level Alignment (SCA) loss to narrow the modality gap while mitigating the effect of noisy pseudo-labels through a soft many-to-many alignment strategy. Extensive experiments on the public SYSU-MM01 and RegDB datasets demonstrate the reliability of the established cross-modality correspondences and the effectiveness of MMM.

Keywords: USL-VI-ReID · Multi-Memory Matching · Noisy Correspondence

1 Introduction

Person re-identification (ReID) is a retrieval task, which aims to match the same person across different cameras, serving critical roles in video surveillance applications like intelligent security [10, 11] and human analysis [22, 23]. However, in low-light conditions, the images captured by visible cameras are far from satisfactory, which renders
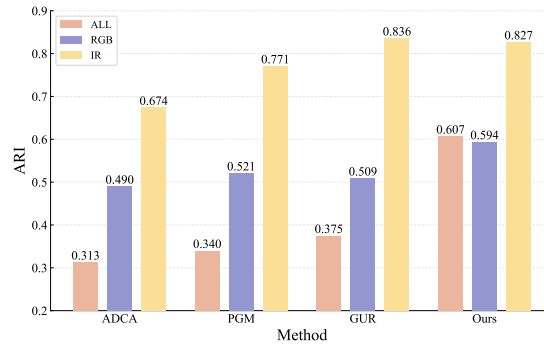

⋆ Corresponding author

Fig. 1: Comparision with different methods on ARI. The ARI indicates the Adjusted Rand Index, which is a similarity measure between two clusterings. The ALL category represents the ARI values of overall pseudo-labels, composed of visible and infrared pseudo-labels, and serves as a metric for evaluating the reliability of cross-modality correspondences.

methods [32, 33, 55, 60] that primarily focus on matching visible images less effective. Fortunately, smart surveillance cameras that can switch from visible to infrared modes in poor lighting environments have become widespread, driving the development of visible-infrared person re-identification (VI-ReID) for the 24-hour surveillance system.

VI-ReID aims at retrieving infrared images of the same person when provided with a visible person image, and vice versa [13,39,48]. Many VI-ReID methods [20,49,56,57] have shown promising progress. However, these methods are based on well-annotated cross-modality data, which is time-consuming and labor-intensive, thereby limiting the practical application of supervised VI-ReID methods in real-world scenarios.

To free the toilsome label process and speed the automation of VI-ReID, several unsupervised VI-ReID (USL-VI-ReID) methods [16, 45, 46, 53] have been proposed, which try to establish cross-modality correspondences by clustering pseudo-labels and have achieved fairly good performance. However, the reliability of pseudo-labels and cross-modality correspondences in USL-VI-ReID still is untouched. We argue the problem is critical to the credibility of USL-VI-ReID. To measure the reliability, we introduce the Adjusted Rand Index (ARI) metric [17], which is a widely recognized metric for clustering evaluation. The larger the ARI value, the better it reflects the degree of overlap between the clustered results and the ground-truth labels. More detailed explanations are presented in **supplementary materials**. In Fig. 1, RGB and IR categories denote the ARI values of visible and infrared pseudo-labels, which can measure the quality of visible and infrared pseudo-labels. Interestingly, we unveil a paradoxical phenomenon: the reliability of cross-modality correspondences in previous methods stands questioned, notwithstanding their demonstrated efficacy, as depicted in Fig. 1 and Tab. 1. This conundrum may arise from the reality that individuals, despite bearing unique identities, manifest overlapping attributes, which tend to merge more closely due to noisy correspondences. Although this amalgamation of similar characteristics can inadvertently heighten the similarity across cross-modality features, which may

lead to increased significant challenges in the precise retrieval of specific persons from a densely populated gallery.

To reduce the noisy cross-modality correspondences in USL-VI-ReID, we develop a novel Multi-Memory Matching (MMM) framework. Multi-memory can store a wider array of distinct characteristics for an identity. For example, Memory 1 can retain front-facing attributes, Memory 2 can capture rear-facing attributes. In short, multi-memory supports a more diverse representation, which is beneficial for the establishment of cross-modality correspondences. Specifically, we propose a Cross-Modality Clustering (CMC) module to generate pseudo-labels. Unlike previous methods, we not only cluster intra-modality samples but also cluster inter-modality samples to learn modality-invariant features. We note that the existing methods typically rely on a single memory to represent individual characteristics and establish cross-modality correspondences. However, a single memory may not capture all individual nuances, including perspective, attire, and other factors, which naturally leads to poor cross-modality correspondences. Therefore, we design a Multi-Memory Learning and Matching (MMLM) module to obtain reliable cross-modality correspondences. We subdivide single memory into multi-memory for a single identity by sub-cluster and compute a cost matrix for multi-memory. To reduce the discrepancy between the two modalities, we propose the Soft Cluster-level Alignment (SCA) loss to narrow the modality gap through soft cluster-level intra- and inter-modality alignment. MMM can achieve fairly good quality of pseudo-labels and cross-modality correspondences compared with several USL-VI-ReID methods, as shown in Fig. 1.

The main contributions are summarized as follows:

- We introduce the ARI metric to evaluate the quality of pseudo-labels and cross-modality correspondences. We observe a curious phenomenon: the cross-modality correspondences of previous methods are not reliable, though they achieve good performance.
- We design a novel Multi-Memory Matching (MMM) framework for unsupervised VI-ReID, which exploits the individual nuances to effectively establish reliable cross-modality correspondences.
- We introduce two effective modules and one loss: Cross-Modality Clustering (CMC), Multi-Memory Learning and Matching (MMLM), and Soft Cluster-level Alignment (SCA). They facilitate the generation of pseudo-labels, establish reliable cross-modality correspondences, and narrow the discrepancy between two modalities while mitigating the influence of noisy pseudo-labels.

2   Related Work

2.1   Supervised Visible-Infrared Person ReID

Visible-infrared person ReID is a challenging cross-modality image retrieval problem [34, 35]. Many works have been proposed to alleviate the large cross-modality gap for VI-ReID, which can be broadly categorized into two classes: image-level alignment and feature-level alignment. The image-level alignment methods [37, 39] try to generate cross-modal images to excavate modality-invariant information. Moreover, several

methods [29, 40, 59] introduce an auxiliary modality to assist the cross-modality retrieval task. The feature-level alignment methods [14, 31, 48, 50, 62] mainly map cross-modal features into a shared feature space to reduce cross-modal differences. For example, SGIEL [9] separates shape-related features from shape-erasure features through orthogonal decomposition to improve the diversity and identification of the learned representations for VI-ReID. However, the above methods heavily rely on large-scale cross-modality data annotation, which is quite expensive and time-consuming.

2.2   Unsupervised Single-Modality Person ReID

Existing unsupervised single-modality person ReID (USL-ReID) methods can be roughly categorized into domain translation-based methods and clustering-based methods. The domain translation-based methods [10–12,27,54] try to transfer the knowledge from the labeled source domain to the unlabeled target domain for USL-ReID. Compared with the former, the clustering-based methods [2,24,36,55] are more challenging, which are trained directly on the unlabeled target domain. The common idea of clustering-based methods is using clustering algorithms [8] to generate pseudo-labels to train a ReID model. Pseudo-labels inevitably contain noise, so it is challenging to assign the correct label to each unlabeled image. Recently, Cluster-Contrast [7] performs contrastive learning at the cluster level with a uni-centroid. However, a uni-centroid cannot represent a cluster well. Therefore, MCRN [42] and DCMIP [61] store multi-centroid representations to completely represent a cluster. Their multi-centroids are obtained through initialization, but these divisions do not accurately represent the real distribution. Although the above methods perform well on USL-ReID, they are not suitable for solving the USL-VI-ReID due to the large cross-modality gap.

2.3   Unsupervised Visible-Infrared Person ReID

The challenge of unsupervised VI-ReID (USL-VI-ReID) is establishing reliable cross-modality correspondence. H2H [21] and OTLA [38] use a well-annotated labeled source domain for pre-training to solve the USL-VI-ReID. Inspired by Cluster-Contrast [7] for USL-ReID, some clustering-based methods [5, 6, 16, 43, 45] are proposed for USL-VI-ReID, they try to establish cross-modality correspondence by clustering pseudo-labels. Recently, it has been shown that the Large-scale Vision-Language Pre-training model, naturally excels in producing textual descriptions for images. To this end, CCLNet [4] leverages the text information from CLIP to improve the USL-VI-ReID task. However, none of the above methods evaluate the reliability of cross-modality correspondence, indeed, their cross-modality correspondence is not reliable. Our method aims to investigate how to establish more reliable cross-modality correspondence for USL-VI-ReID.

3   Methodology

The framework of MMM is illustrated in Fig. 2. We begin by employing the Cross-Modality Clustering (CMC) module to generate pseudo-labels. Building upon CMC,
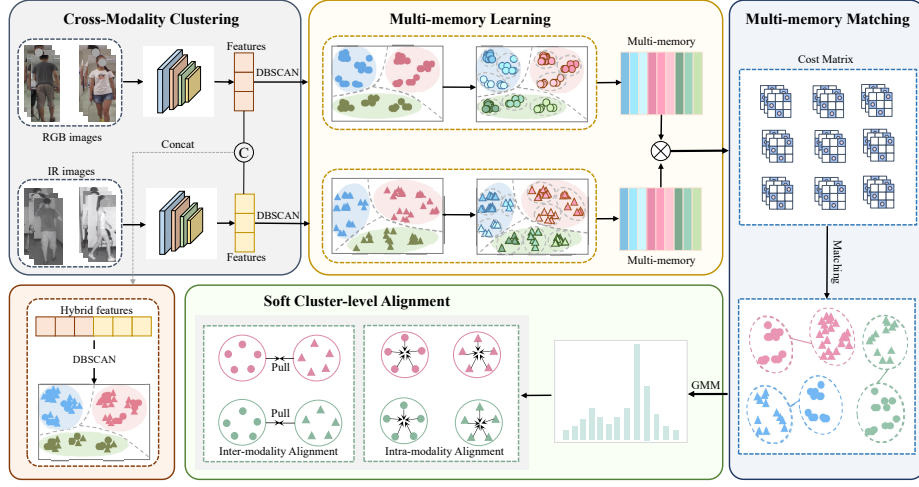
Fig. 2: The pipeline of MMM. Different colors indicate different persons, $\bigcirc$ and $\triangle$ indicate visible and infrared features. It contains the Cross-Modality Clustering module (Baseline, described in Sec. 3.2) and two key novel components: Multi-Memory Learning and Matching (MMLM, described in Sec. 3.3) and Soft Cluster-level Alignment (SCA, described in Sec. 3.4).

we propose a novel Multi-Memory Learning and Matching (MMLM) module to effectively establish cross-modality correspondences. Finally, we propose the Soft Cluster-level Alignment (SCA) loss to narrow the gap between two modalities while mitigating the impact of noisy pseudo-labels through soft cluster-level intra- and inter-modality alignment.

3.1   Notation Definition

Suppose we have a USL-VI-ReID dataset denoted as $D = \{V, R\}$. Here, $V = \{v_i\}_{i=1}^{N}$ represents the visible images with N samples, and $R = \{r_i\}_{i=1}^{M}$ denotes the infrared images with M samples. We initialize their pseudo-labels as Y^t, where $t \in \{v, r\}$. Let N_p and M_p represent the number of visible and infrared samples with ID p, where $p \in \{1, 2, ..., P^t\}$ and P^t is the total number of person identities for modality t. The respective feature sets of these images are denoted as $F^v = \{f_1^v, f_2^v, \ldots, f_N^v\}$ for visible samples and $F^r = \{f_1^r, f_2^r, \ldots, f_M^r\}$ for infrared samples, respectively. Our goal is to develop a cross-modality person ReID model without utilizing any labels.

3.2   Cross-Modality Clustering

Most USL-VI-ReID methods typically use clustering algorithms to generate pseudo-labels. Following this paradigm, we employ the DBSCAN algorithm [8] to generate pseudo-labels for all images, as described:

$$Y^t = DBSCAN(F^t). \tag{1}$$

Unlike previous methods, we not only cluster intra-modality samples ($t = v$ or $t =$ r) but also cluster inter-modality samples ($t = \{v, r\}$) to indirectly build cross-modality correspondence.

At the beginning of every training iteration, we calculate and store the memory for each cluster as follows:

$$\boldsymbol{C}_{V^p} = \frac{1}{N_p} \sum_{i=1}^{N_p} f(V_i^p), \tag{2}$$

$$\boldsymbol{C}_{R^p} = \frac{1}{M_p} \sum_{i=1}^{M_p} f(R_i^p), \tag{3}$$

$$\boldsymbol{C}_{VR^p} = \frac{1}{A_p} \sum_{i=1}^{A_p} f(VR_i^p), \tag{4}$$

where $f(\cdot)$ is a function designated for extracting features from images across diverse modalities. We use superscripts to denote specified identity, V^p and R^p denote the visible and infrared modality of the same identity sample sets with ID p, respectively. VR^p represents the combined set of both modalities with A_p samples of the same ID p.

Then, we optimize the feature extractor using ClusterNCE [7] loss, computed as:

$$L_V = -\log \frac{\exp\left(C_V^+ \cdot F^v / \tau\right)}{\sum_{p=1}^{P^v} \exp\left(C_{V^p} \cdot F^v / \tau\right)}, \tag{5}$$

$$L_R = -\log \frac{\exp\left(C_R^+ \cdot F^r / \tau\right)}{\sum_{p=1}^{P^r} \exp\left(C_{R^p} \cdot F^r / \tau\right)}, \tag{6}$$

$$L_{VR} = -\log \frac{\exp\left(C_{VR}^+ \cdot [F^v, F^r] / \tau\right)}{\sum_{p=1}^{P^{v,r}} \exp\left(C_{VR^p} \cdot [F^v, F^r] / \tau\right)}, \tag{7}$$

where C^+ is the positive memory representation and the τ is a temperature hyper-parameter.

The CMC loss is defined as:

$$L_{CMC} = L_V + L_R + L_{VR}. \tag{8}$$

3.3 Multi-Memory Learning and Matching

The CMC optimizes the feature extractor using a single memory, but a single memory may not fully capture individual nuances, such as perspective and attire. Moreover, the CMC does not directly establish relations between the two modalities, thereby limiting its effectiveness in cases with significant modality discrepancies. To more effectively capture individual nuances and bridge the gap between the visible and infrared modalities, we propose the Multi-Memory Learning and Matching (MMLM) module, which mines a holistic representation and establishes reliable cross-modality correspondences.

Specifically, we further subdivide single memory into multi-memory for a single identity, which can be formulated as a sub-cluster:

$$\min_{F_{C_{V_i^p}}} \{\sum_{i=1}^{n}\{\left\|f^v - K_{C_{V_i^p}}\right\|_2^2, \forall f^v \in F_{C_{V_i^p}}\}\}, \tag{9}$$

$$\min_{F_{C_{R_i^p}}} \{\sum_{i=1}^{n}\{\left\|f^r - K_{C_{R_i^p}}\right\|_2^2, \forall f^r \in F_{C_{R_i^p}}\}\}, \tag{10}$$

where $F_{C_{V_i^p}}$ and $F_{C_{R_i^p}}$ represent the i-th visible and infrared feature sets of ID p, respectively. n is the number of memories for a single identity.

$$K_{C_{V_i^p}} = \frac{1}{|F_{C_{V_i^p}}|} \sum_{f^v \in F_{C_{V_i^p}}} f^v, \tag{11}$$

$$K_{C_{R_i^p}} = \frac{1}{|F_{C_{R_i^p}}|} \sum_{f^r \in F_{C_{R_i^p}}} f^r, \tag{12}$$

where $K_{C_{V^p}}$ and $K_{C_{R^p}}$ represent the visible and infrared multi-memory of ID p.

By employing the multi-memory learning strategy, we achieve more diverse memories for a single identity. However, these memories still exhibit a strong implicit correlation with the modality, which negatively impacts the establishment of cross-modality correspondences. Inspired by PGM [43], we transform the cross-modality multi-memory matching problem into a weighted bipartite graph matching. The goal is to match each visible cluster with the corresponding identity infrared cluster while minimizing the cost, which is formulated as follows:

$$\min_{Q} M^T Q$$
$$\text{s.t. } \forall p \in [P^v], \forall p' \in [P^r] : Q_p^{p'} \in \{0, 1\},$$
$$\forall p \in [P^v] : \sum_{p' \in [P^r]} Q_p^{p'} \leq 1, \tag{13}$$
$$\forall p' \in [P^r] : \sum_{p \in [P^v]} Q_p^{p'} = 1,$$

where $Q = \left\{Q_p^{p'}\right\} \in \mathbb{R}^{P^v \times P^r \times 1}$ indicates whether K_{V^p} and $K_{R^{p'}}$ belong to the same person $\left(Q_p^{p'} = 1\right)$ or not $\left(Q_p^{p'} = 0\right)$. M and $[P^t]$ denote cost matrix and $\{1, \ldots, P^t\}$, respectively. We design a simple yet effective cost expression for cross-modality multi-memory matching as follows:

$$M(K_{C_{V^p}}, K_{C_{R^{p'}}}) = \sum_{i=1}^{n} \min_{j \in \{1, \cdots, n\}} \|K_{V_i^p}, K_{R_j^{p'}}\|_2, \tag{14}$$

Finally, we transfer the infrared pseudo-labels to the visible pseudo-labels, and the visible pseudo-labels are updated by:

$$Y^v := QY^r. \tag{15}$$

3.4   Soft Cluster-level Alignment

Pseudo-labels inherently contain noise, a problem that is not exempt even in human annotations [48], leading to a reduction in performance. The method [1] illustrated that deep neural networks initially learn from simple samples before accommodating noisy labels. Building on this insight, we assess the confidence associated with each label. To do so, we employ a two-component Gaussian Mixture Model (GMM) to model the loss distribution:

$$L_{ID}^v = -\log p\left(Y^v \mid C\left(F^v\right)\right), \tag{16}$$

$$p(L_{ID}^v \mid \theta) = \sum_{k=1}^{2} \pi_k \phi(L_{ID}^v \mid k), \tag{17}$$

where $C(\cdot)$ acts as an identity classifier. π_k represents the mixture coefficient, while $\phi(L_{ID}^v \mid k)$ denotes the probability density of the k-th component.

Subsequently, the confidence is determined by computing its posterior probability, detailed as:

$$W^v = p\left(k \mid L_{ID}^v\right), \tag{18}$$

where k refers to the Gaussian component with a smaller mean, while $p\left(k \mid L_{ID}^v\right)$ indicates the responsiveness of L_{ID}^v to the k-th component. In the same way, we can obtain the confidence W^r and W^{vr}.

To penalize the noise during optimization, the memories in Eq. (2), (3), (4) are updated by:

$$C_{V^p} := \frac{1}{N_p} \sum_{i=1}^{N_p} f(V_i^p) W_{V_i^p}, \tag{19}$$

$$C_{R^p} := \frac{1}{M_p} \sum_{i=1}^{M_p} f(R_i^p) W_{R_i^p}, \tag{20}$$

$$C_{VR^p} := \frac{1}{A_p} \sum_{i=1}^{A_p} f(VR_i^p) W_{VR_i^p}, \tag{21}$$

where $W_{V_i^p}$, $W_{R_i^p}$, and $W_{VR_i^p}$ denote the confidences of samples V_i^p, R_i^p, and VR_i^p, respectively.

To reduce the intra-modality discrepancy, we employ the distilled C_{V^p} and C_{R^p} to align every sample of ID p to its corresponding memory in each modality. The cluster-level intra-modality alignment loss L_{Intra} is proposed as:

$$
\begin{aligned}
L_{Intra} &= L_{Intra}^V + L_{Intra}^R \\
&= \sum_{p=1}^{P^v} \sum_{f^v \in F_p^v} \|f^v - C_{V^p}\|_2^2 \\
&+ \sum_{p=1}^{P^r} \sum_{f^r \in F_p^r} \|f^r - C_{R^p}\|_2^2,
\end{aligned} \tag{22}
$$

where F_p^v, F_p^r denote visible feature and infrared feature sets of ID p, respectively.

Since VI-ReID is a many-to-many matching problem, we propose cluster-level inter-modality alignment loss, which forces the feature distribution of the samples from the visible modality to be similar to the feature distribution of the samples from the infrared modality and vice versa by:

$$
\begin{aligned}
L_{Inter} &= L_{Inter}^V + L_{Inter}^R \\
&= \frac{1}{P} \sum_{p=1}^{P} (\frac{1}{2} D(F_p^v, sg(F_p^r)) \\
&\quad + \frac{1}{2} D(F_p^r, sg(F_p^v))),
\end{aligned}
\tag{23}
$$

where $sg(\cdot)$ represents the stop-gradient operation, and $D(i, j)$ represents the distance between distributions i and j. P is $\min(P^v, P^r)$. In this paper, we employ the squared Maximum Mean Discrepancy (MMD^2) [15] to quantify the discrepancy between distributions. MMD^2 is a commonly used non-parametric metric in domain adaptation and has been observed to outperform other metrics, such as KL divergence in empirical studies, MMD^2 is constructed as:

$$
\begin{aligned}
\text{MMD}^2(F_p^r, F_p^v) &= \frac{1}{|F_p^r|^2} \sum_{f_i^r \in F_p^r} \sum_{f_j^r \in F_p^r} z(f_i^r, f_j^r) \\
&\quad + \frac{1}{|F_p^v|^2} \sum_{f_i^v \in F_p^v} \sum_{f_i^v \in F_p^v} z(f_i^v, f_j^v) \\
&\quad - \frac{2}{|F_p^r||F_p^v|} \sum_{f_i^r \in F_p^r} \sum_{f_j^v \in F_p^r} z(f_i^r, f_j^v),
\end{aligned}
\tag{24}
$$

where $z(s, s') = \exp(\frac{-\|s - s'\|_2^2}{2\sigma^2})$ is a Gaussian kernel.

The SCA loss is defined as:

$$
L_{SCA} = \lambda_{Intra} L_{Intra} + \lambda_{Inter} L_{Inter},
\tag{25}
$$

where λ_{Intra} and λ_{Inter} are the balancing weights.

Overall Loss. The total loss for training the model is defined by the following equation:

$$
L_{overall} = L_{CMC} + L_{SCA}.
\tag{26}
$$

4   Experiments

In this section, we conduct comprehensive experiments to verify the effectiveness of MMM. First, we compare MMM with several state-of-the-art methods under three settings, *i.e.*, supervised visible-infrared person ReID (SVI-ReID), semi-supervised visible-infrared person ReID (SSVI-ReID) and unsupervised visible-infrared person ReID (USL-VI-ReID). After that, we perform ablation studies to evaluate the effectiveness of each

Table 1: Comparisons with state-of-the-art methods on SYSU-MM01 and RegDB, *i.e.*, supervised visible-infrared person ReID (SVI-ReID), semi-supervised visible-infrared person ReID (SSVI-ReID) and unsupervised visible-infrared person ReID (USL-VI-ReID). All methods are measured by Rank-1 (%) and mAP (%). GUR* denotes the results without camera information.

Settings			SYSU-MM01				RegDB			
			All Search		Indoor Search		Visible2Thermal		Thermal2Visible	
Type	Method	Venue	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
SVI-ReID	AGW [51]	TRAMI'21	47.5	47.7	54.2	63.0	70.1	66.4	70.5	65.9
	NFS [3]	CVPR'21	56.9	55.5	62.8	69.8	80.5	72.1	78.0	69.8
	LbA [28]	ICCV'21	55.4	54.1	58.5	66.3	74.2	67.6	72.4	65.5
	CAJ [50]	ICCV'21	69.9	66.9	76.3	80.4	85.0	79.1	84.8	77.8
	DART [48]	CVPR'22	68.7	66.3	72.5	78.2	83.6	75.7	82.0	73.8
	DEEN [58]	CVPR'23	74.7	71.8	80.3	83.3	91.1	85.1	89.5	83.4
	PartMix [19]	CVPR'23	77.8	74.6	81.5	84.4	85.7	82.3	84.9	82.5
SSVI-ReID	OTLA [38]	ECCV'22	48.2	43.9	47.4	56.8	49.9	41.8	49.6	42.8
	TAA [44]	TIP'23	48.8	42.3	50.1	56.0	62.2	56.0	63.8	56.5
	DPIS [30]	ICCV'23	58.4	55.6	63.0	70.0	62.3	53.2	61.5	52.7
USL-VI-ReID	OTLA [38]	ECCV'22	29.9	27.1	29.8	38.8	32.9	29.7	32.1	28.6
	ADCA [47]	MM'22	45.5	42.7	50.6	59.1	67.2	64.1	68.5	63.8
	ADCA+MMM	-	49.7	44.7	56.2	62.5	77.8	70.9	76.5	69.1
	NGLR [5]	MM'23	50.4	47.4	53.5	61.7	85.6	76.7	82.9	75.0
	MBCCM [16]	MM'23	53.1	48.2	55.2	62.0	83.8	77.9	82.8	76.7
	CCLNet [4]	MM'23	54.0	50.2	56.7	65.1	69.9	65.5	70.2	66.7
	PGM [43]	CVPR'23	57.3	51.8	56.2	62.7	69.5	65.4	69.9	65.2
	CHCR [26]	TCSVT'23	59.5	59.1	-	-	69.3	64.7	70.0	65.9
	GUR* [45]	ICCV'23	61.0	57.0	64.2	69.5	73.9	70.2	75.0	69.9
	PCLHD [18]	arXiv'24	64.4	58.7	69.5	74.4	84.3	80.7	82.7	78.4
	MMM	-	61.6	57.9	64.4	70.4	**89.7**	80.5	85.8	77.0
	MMM+PCLHD	**-**	**65.9**	**61.8**	**70.3**	**74.9**	89.6	**83.7**	**87.0**	**80.9**

module in MMM. Finally, we perform a discussion and analysis of the hyper-parameters and visualization. If not specified, we conduct analysis experiments on SYSU-MM01 in the single-shot & all-search mode.

4.1 Experimental Setting

Dataset. We evaluate MMM on two benchmarks, *i.e.*, **SYSU-MM01** [41] and **RegDB** [25]. SYSU-MM01 is a large-scale visible-infrared person ReID dataset, which is collected from four visible cameras and two infrared cameras in both indoor and outdoor scenes. RegDB is a relatively small dataset, which is collected by one visible and one infrared camera in a dual-camera system.

Evaluation Protocols. Cumulative Matching Characteristics [52] and mean Average Precision (mAP) are adopted as the evaluation metrics on two datasets to evaluate the performance of MMM quantitatively. For fair comparisons, we report the results of all-search mode and indoor-search mode with the official code on SYSU-MM01. Following [50], We also report the results on RegDB by randomly splitting the training and testing set 10 times in visible-to-thermal and thermal-to-visible modes.

4.2   Implementation Details

We adopt ResNet50, which is initialized with the ImageNet pre-trained weights, as the shared backbone to extract 2048d features. MMM is implemented in PyTorch. The total number of training epochs is 80. At each training step, we randomly sample 8 IDs, of which 4 visible and 4 infrared images are chosen to formulate a batch. Training images are resized to 288×144 and random horizontal flipping and random crop are used for data augmentation [50]. SGD optimizer is adopted to train the model with the momentum setting to 0.9 and weight decay setting to $5e - 4$. The Intra module is added from the 1st epoch and the Inter module is added from the 15th epoch. The loss temperature τ is set to 0.05. The hyperparameters 'eps' and 'min_samples' in DBSCAN are set to 0.6 and 4.

Table 2: Ablation studies on SYSU-MM01 in all search mode and indoor search mode. "Baseline" means the model trained only with the CMC module. Rank-R accuracy(%) and mAP(%) are reported.

	Method				All Search					Indoor Search				
Order	Baseline	MMLM	Intra	Inter	Rank-1	Rank-5	Rank-10	Rank-20	mAP	Rank-1	Rank-5	Rank-10	Rank-20	mAP
1	✓				51.74	78.67	87.87	94.76	49.81	56.34	84.66	92.77	96.98	64.46
2	✓	✓			55.15	81.65	90.53	96.46	52.21	58.76	85.21	93.06	97.16	65.47
3	✓	✓	✓		58.48	83.69	91.79	97.15	55.05	62.19	86.95	93.60	97.64	68.09
4	✓	✓		✓	57.26	82.34	90.84	96.93	53.81	60.26	85.77	93.16	97.36	66.66
5	✓	✓	✓	✓	**61.56**	**85.66**	**93.33**	**98.03**	**57.92**	**64.37**	**88.80**	**95.01**	**98.20**	**70.40**

4.3   Results and Analysis

To clearly demonstrate the effectiveness of MMM, we compare MMM with several state-of-the-art methods under three settings, *i.e.*, SVI-ReID, SSVI-ReID, and USL-VI-ReID. The quantitative results on SYSU-MM01 and RegDB are shown in Tab. 1.
Comparison with SSVI-ReID Methods. We compared MMM with three state-of-the-art SSVI-ReID methods. Notably, MMM not only outpaced these methods but did so without relying on any form of annotations. This stands in stark contrast to the SSVI-ReID methods, which rely on annotations of visible images to achieve their results.
Comparison with USL-VI-ReID Methods. Compared with eight state-of-the-art USVI-ReID methods, MMM consistently performs better than existing USL-VI-ReID methods by a significant margin. PCLHD [18] is proposed to learn more discriminative cross-modality features, and our method with PCLHD can achieve 65.9% in Rank-1 and 61.8% in mAP. ADCA [47] with MMM also achieve consistently improved performance. Moreover, the results are surprising on RegDB, MMM improves the Rank-1 and mAP accuracy by a large margin of 15.8% and 10.3% compared to GUR under visible to thermal mode.
Comparison with SVI-ReID Methods. Surprisingly, MMM performs better than several supervised methods, including AGW [51], NFS [3], and LbA [28]. The results show the effectiveness of MMM. However, we have to acknowledge that there is still a certain
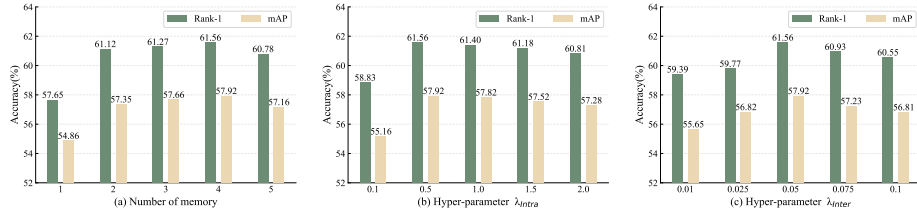
Fig. 3: The effect of hyper-parameter n, λ_{Intra} and λ_{Inter} with different values on SYSU-MM01.

gap between MMM and many SVI-ReID methods due to the absence of cross-modality data annotations.

The above results clearly show that MMM is effective, which highlights the significant potential of MMM in addressing USL-VI-ReID challenges.

4.4   Ablation Study

To further analyze the effectiveness of the Multi-Memory Learning and Matching (MMLM), the Soft Cluster-level Alignment (SCA), we conduct ablation studies on SYSU-MM01 under both all-search and indoor-search modes. The results are reported in Tab. 2.
Baseline. Order 1 denotes that the model is trained only with the CMC module. Although it achieves a promising performance on SYSU-MM01, it does not directly establish relations between the two modalities, which limits the performance.
Effective of MMLM. The effectiveness of the MMLM module is revealed by comparing Order 1 and Order 2. The MMLM improves 3.41% in Rank-1 and 2.40% in mAP on SYSU-MM01. The results, combined with Fig. 1, demonstrate that the MMLM can help align visible and infrared pseudo-labels to establish cross-modality correspondences.
Effective of Intra in SCA. As shown in Order 3 of Tab. 2, the performance is improved to 58.48% in Rank-1 and 55.05% in mAP when adding the cluster-level intra-modality loss (Intra) in SCA, which shows the effectiveness of Intra in reducing the discrepancy of intra-modality.
Effective of Inter in SCA. The cluster-level inter-modality alignment loss (Inter) is proposed to reduce the discrepancy of inter-modality, MMM can reach 57.26% in Rank-1 and 53.81% in mAP when adding it. Moreover, when combining Inter with Intra, MMM achieves the best performance with 61.56% in Rank-1 and 57.92% in mAP, which surpasses the baseline by a large margin of 9.82% in Rank-1 and 8.11% in mAP.

The above results show that cluster-level intra- and inter-modality alignment loss can complement each other, which proves the effectiveness of the SCA loss.

4.5   Analysis of Hyper-parameters

We analyze the key hyper-parameters of MMM on SYSU-MM01, *i.e.*, the number of memories n, λ_{Intra} and λ_{Inter}. In Fig. 3 (a), we vary the number of memories from 1 to 5 while keeping the λ_{Intra} and λ_{Inter} fixed, which shows MMM achieves the best
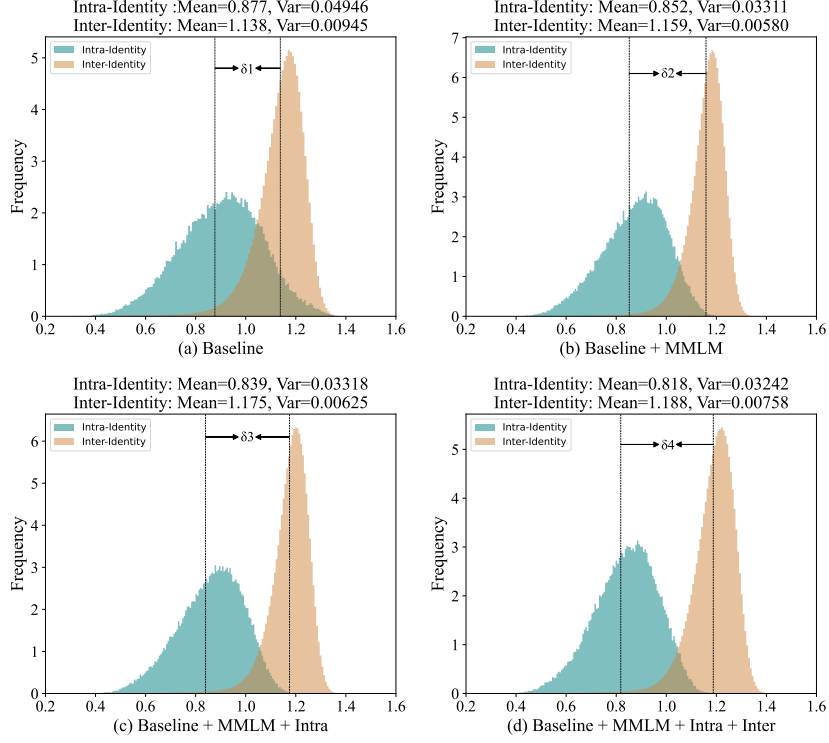
Fig. 4: The intra-identity and inter-identity distances on SYSU-MM01, where δ_i denotes the gap between the intra-identity distance mean and the inter-identity distance mean.

performance with 61.56% in Rank-1 and 57.92% in mAP when $n = 4$. Moreover, to balance the contribution between the cluster-level intra- and inter-modality alignment loss in SCA, we study the effect of λ_{Intra} and λ_{Inter} by fixing one and adjusting the other. To be specific, we maintain the $\lambda_{Inter} = 0.05$ and tune the value of λ_{Intra} in $[0.1, 0.5, 1.0, 1.5, 2.0]$ (Fig. 3 (b)), while fix the $\lambda_{Intra} = 0.5$ and explore the λ_{Inter} on different values which vary in $[0.01, 0.025, 0.05, 0.075, 0.1]$ (Fig. 3 (c)). We can observe that MMM achieves high accuracy under different combinations with λ_{Intra} and λ_{Inter}, which shows the performance of MMM is not sensitive to λ_{Intra} and λ_{Inter}, and the best performance is achieved with $\lambda_{Intra} = 0.5$ and $\lambda_{Inter} = 0.05$.

4.6 Qualitative Analysis

To further illustrate the effectiveness of MMM, we visualize the intra-identity and inter-identity distances on SYSU-MM01 in Fig. 4. As shown in Fig. 4 (a)-(d), with the addition of the proposed methods, the means of intra-identity distances gradually decrease while the means of inter-identity distances gradually increase, which makes the intra-identity and inter-identity features distributions are pushed away ($\delta_1 < \delta_2 < \delta_3 < \delta_4$).
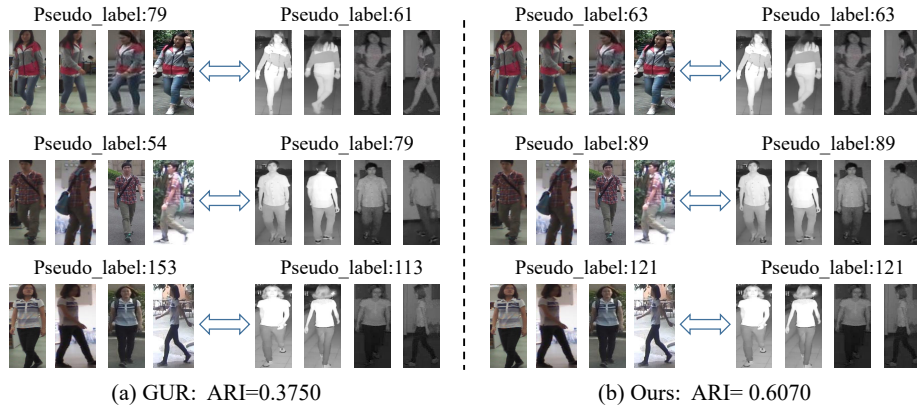
Pseudo_label:79   Pseudo_label:61   Pseudo_label:63   Pseudo_label:63

Pseudo_label:54   Pseudo_label:79   Pseudo_label:89   Pseudo_label:89

Pseudo_label:153   Pseudo_label:113   Pseudo_label:121   Pseudo_label:121

(a) GUR: ARI=0.3750   (b) Ours: ARI= 0.6070

Fig. 5: The Visualization of the pseudo-labels of the same identity with different modalities.

The results show that MMM can effectively reduce the cross-modality distances between the same identity samples and push the distance between different identity samples far away.

Moreover, we also visualize the pseudo-labels of the same identity with different modalities, where we randomly choose 3 person identities, where each identity consists of 4 visible images and infrared images. As shown in Fig. 5, persons of the same identity in different modalities have the same pseudo-label in MMM (right) compared with GUR (left), which shows that MMM can establish more reliable cross-modality correspondences.

5   Conclusion

In this paper, we introduce a metric, the Adjusted Rand Index, to measure cross-modality correspondences and clustered pseudo-labels, exploring the establishment of reliable cross-modality correspondences for USL-VI-ReID. To this end, we propose a Multi-Memory Matching (MMM) framework. Firstly, we design a Cross-Modality Clustering (CMC) module to generate pseudo-labels. Instead of previous methods, we employ multi-memory in the Multi-Memory Learning and Matching (MMLM) module to capture individual nuances and establish reliable cross-modality correspondences. Additionally, we present a Soft Cluster-level Alignment (SCA) loss to reduce the cross-modality gap while mitigating the effect of noisy pseudo-labels. Comprehensive experimental results show that MMM can establish reliable cross-modality correspondences and outperforms existing USL-VI-ReID methods on SYSU-MM01 and RegDB.

References

1. Arpit, D., Jastrzebski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M.S., Maharaj, T., Fischer, A., Courville, A.C., Bengio, Y., Lacoste-Julien, S.: A closer look at memorization in deep networks. In: ICML. pp. 233–242 (2017) 8

2. Chen, H., Lagadec, B., Brémond, F.: ICE: inter-instance contrastive encoding for unsupervised person re-identification. In: ICCV. pp. 14940–14949 (2021) 4

3. Chen, Y., Wan, L., Li, Z., Jing, Q., Sun, Z.: Neural feature search for rgb-infrared person re-identification. In: CVPR. pp. 587–597 (2021) 10, 11

4. Chen, Z., Zhang, Z., Tan, X., Qu, Y., Xie, Y.: Unveiling the power of clip in unsupervised visible-infrared person re-identification. In: ACM MM. pp. 3667–3675 (2023) 4, 10

5. Cheng, D., Huang, X., Wang, N., He, L., Li, Z., Gao, X.: Unsupervised visible-infrared person reid by collaborative learning with neighbor-guided label refinement. ArXiv:2305.12711 (2023) 4, 10

6. Cho, Y., Kim, W.J., Hong, S., Yoon, S.: Part-based pseudo label refinement for unsupervised person re-identification. In: CVPR. pp. 7298–7308 (2022) 4

7. Dai, Z., Wang, G., Yuan, W., Zhu, S., Tan, P.: Cluster contrast for unsupervised person re-identification. In: ACCV. pp. 319–337 (2022) 4, 6

8. Ester, M., Kriegel, H., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: KDD. pp. 226–231 (1996) 4, 5

9. Feng, J., Wu, A., Zheng, W.: Shape-erased feature learning for visible-infrared person re-identification. In: CVPR. pp. 22752–22761 (2023) 4

10. Fu, Y., Wei, Y., Wang, G., Zhou, Y., Shi, H., Huang, T.S.: Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In: ICCV. pp. 6111–6120 (2019) 1, 4

11. Ge, Y., Chen, D., Li, H.: Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. In: ICLR (2020) 1, 4

12. Ge, Y., Zhu, F., Chen, D., Zhao, R., Li, H.: Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. In: NeurIPS (2020) 4

13. Gong, Y., Huang, L., Chen, L.: Person re-identification method based on color attack and joint defence. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 4312–4321. IEEE (2022) 2

14. Gong, Y., Zhong, Z., Luo, Z., Qu, Y., Ji, R., Jiang, M.: Cross-modality perturbation synergy attack for person re-identification. CoRR **abs/2401.10090** (2024) 4

15. Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A.: A kernel two-sample test. The Journal of Machine Learning Research **13**(1), 723–773 (2012) 9

16. He, L., Wang, N., Zhang, S., Wang, Z., Gao, X., et al.: Efficient bilateral cross-modality cluster matching for unsupervised visible-infrared person reid. ArXiv:2305.12673 (2023) 2, 4, 10

17. Hubert, L., Arabie, P.: Comparing partitions. Journal of classification **2**, 193–218 (1985) 2

18. Jiangming Shi, Xiangbo Yin, Y.W.X.L.Y.X.Y.Q.: Progressive contrastive learning with multi-prototype for unsupervised visible-infrared person re-identification. arXiv:2402.19026 (2024) 10, 11

19. Kim, M., Kim, S., Park, J., Park, S., Sohn, K.: Partmix: Regularization strategy to learn part discovery for visible-infrared person re-identification. In: CVPR. pp. 18621–18632 (2023) 10

20. Li, H., Ye, M., Zhang, M., Du, B.: All in one framework for multimodal re-identification in the wild. In: CVPR. pp. 17459–17469 (2024) 2

21. Liang, W., Wang, G., Lai, J., Xie, X.: Homogeneous-to-heterogeneous: Unsupervised learning for rgb-infrared person re-identification. IEEE Trans. Image Process. **30**, 6392–6407 (2021) 4

22. Lin, L., Liu, H., Liang, J., Li, Z., Feng, J., Han, H.: Consensus-agent deep reinforcement learning for face aging. IEEE Transactions on Image Processing (2024) 1

23. Lin, L., Wang, T., Liu, H., Zhu, C., Chen, J.: Toward quantifiable face age transformation under attribute unbias. IEEE Transactions on Circuits and Systems for Video Technology (2024) 1

24. Lin, Y., Xie, L., Wu, Y., Yan, C., Tian, Q.: Unsupervised person re-identification via softened similarity learning. In: CVPR. pp. 3387–3396 (2020) 4

25. Nguyen, D.T., Hong, H.G., Kim, K., Park, K.R.: Person recognition system based on a combination of body images from visible light and thermal cameras. Sensors **17**(3), 605 (2017) 10

26. Pang, Z., Wang, C., Zhao, L., Liu, Y., Sharma, G.: Cross-modality hierarchical clustering and refinement for unsupervised visible-infrared person re-identification. IEEE Transactions on Circuits and Systems for Video Technology pp. 1–1 (2023) 10

27. Pang, Z., Zhao, L., Liu, Q., Wang, C.: Camera invariant feature learning for unsupervised person re-identification. IEEE transactions on multimedia **25**, 6171–6182 (2022) 4

28. Park, H., Lee, S., Lee, J., Ham, B.: Learning by aligning: Visible-infrared person re-identification using cross-modal correspondences. In: ICCV. pp. 12026–12035 (2021) 10, 11

29. Shi, J., Yin, X., Zhang, D., Qu, Y.: Visible embraces infrared: Cross-modality person re-identification with single-modality supervision. In: 2023 China Automation Congress (CAC). pp. 4781–4787. IEEE (2023) 4

30. Shi, J., Zhang, Y., Yin, X., Xie, Y., Zhang, Z., Fan, J., Shi, Z., Qu, Y.: Dual pseudo-labels interactive self-training for semi-supervised visible-infrared person re-identification. In: ICCV. pp. 11218–11228 (2023) 10

31. Sun, H., Liu, J., Zhang, Z., Wang, C., Qu, Y., Xie, Y., Ma, L.: Not all pixels are matched: Dense contrastive learning for cross-modality person re-identification. In: ACM MM. pp. 5333–5341 (2022) 4

32. Tan, L., Dai, P., Ji, R., Wu, Y.: Dynamic prototype mask for occluded person re-identification. In: ACM MM. pp. 531–540 (2022) 2

33. Tan, L., Xia, J., Liu, W., Dai, P., Wu, Y., Cao, L.: Occluded person re-identification via saliency-guided patch transfer. In: AAAI. vol. 38, pp. 5070–5078 (2024) 2

34. Tang, Y., Yu, J., Gai, K., Wang, Y., Hu, Y., Xiong, G., Wu, Q.: Align before search: Aligning ads image to text for accurate cross-modal sponsored search (2023) 3

35. Tang, Y., Yu, J., Gai, K., Zhuang, J., Xiong, G., Hu, Y., Wu, Q.: Context-i2w: Mapping images to context-dependent words for accurate zero-shot composed image retrieval. In: AAAI. vol. 38, pp. 5180–5188 (2024) 3

36. Wang, D., Zhang, S.: Unsupervised person re-identification via multi-label classification. In: CVPR. pp. 10978–10987 (2020) 4

37. Wang, G., Yang, Y., Zhang, T., Cheng, J., Hou, Z., Tiwari, P., Pandey, H.M.: Cross-modality paired-images generation and augmentation for rgb-infrared person re-identification. Neural Networks **128**, 294–304 (2020) 3

38. Wang, J., Zhang, Z., Chen, M., Zhang, Y., Wang, C., Sheng, B., Qu, Y., Xie, Y.: Optimal transport for label-efficient visible-infrared person re-identification. In: ECCV. pp. 93–109 (2022) 4, 10

39. Wang, Y., Liu, X., Zhang, P., Lu, H., Tu, Z., Lu, H.: Top-reid: Multi-spectral object re-identification with token permutation. In: AAAI. vol. 38, pp. 5758–5766 (2024) 2, 3

40. Wei, Z., Yang, X., Wang, N., Gao, X.: Syncretic modality collaborative learning for visible infrared person re-identification. In: ICCV. pp. 225–234 (2021) 4

41. Wu, A., Zheng, W., Yu, H., Gong, S., Lai, J.: Rgb-infrared cross-modality person re-identification. In: ICCV. pp. 5390–5399 (2017) 10

42. Wu, Y., Huang, T., Yao, H., Zhang, C., Shao, Y., Han, C., Gao, C., Sang, N.: Multi-centroid representation network for domain adaptive person re-id. In: AAAI. pp. 2750–2758 (2022) 4

43. Wu, Z., Ye, M.: Unsupervised visible-infrared person re-identification via progressive graph matching and alternate learning. In: CVPR. pp. 9548–9558 (2023) 4, 7, 10

44. Yang, B., Chen, J., Ma, X., Ye, M.: Translation, association and augmentation: Learning cross-modality re-identification from single-modality annotation. IEEE Transactions on Image Processing **32**, 5099–5113 (2023) 10

45. Yang, B., Chen, J., Ye, M.: Towards grand unified representation learning for unsupervised visible-infrared person re-identification. In: ICCV. pp. 11069–11079 (2023) 2, 4, 10

46. Yang, B., Chen, J., Ye, M.: Shallow-deep collaborative learning for unsupervised visible-infrared person re-identification. In: CVPR. pp. 16870–16879 (2024) 2

47. Yang, B., Ye, M., Chen, J., Wu, Z.: Augmented dual-contrastive aggregation learning for unsupervised visible-infrared person re-identification. In: ACM MM. pp. 2843–2851 (2022) 10, 11

48. Yang, M., Huang, Z., Hu, P., Li, T., Lv, J., Peng, X.: Learning with twin noisy labels for visible-infrared person re-identification. In: CVPR. pp. 14288–14297 (2022) 2, 4, 8, 10

49. Yang, M., Huang, Z., Peng, X.: Robust object re-identification with coupled noisy labels. IJCV pp. 1–19 (2024) 2

50. Ye, M., Ruan, W., Du, B., Shou, M.Z.: Channel augmented joint learning for visible-infrared recognition. In: ICCV. pp. 13547–13556 (2021) 4, 10, 11

51. Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., Hoi, S.C.H.: Deep learning for person re-identification: A survey and outlook. IEEE Trans. Pattern Anal. Mach. Intell. pp. 2872–2893 (2022) 10, 11

52. Ye, M., Wang, Z., Lan, X., Yuen, P.C.: Visible thermal person re-identification via dual-constrained top-ranking. In: IJCAI. pp. 1092–1099 (2018) 10

53. Yin, X., Shi, J., Zhang, Y., Lu, Y., Zhang, Z., Xie, Y., Qu, Y.: Robust pseudo-label learning with neighbor relation for unsupervised visible-infrared person re-identification. CoRR **abs/2405.05613** (2024) 2

54. Zhai, Y., Ye, Q., Lu, S., Jia, M., Ji, R., Tian, Y.: Multiple expert brainstorming for domain adaptive person re-identification. In: ECCV. vol. 12352, pp. 594–611 (2020) 4

55. Zhang, G., Zhang, H., Lin, W., Chandran, A.K., Jing, X.: Camera contrast learning for unsupervised person re-identification. IEEE Trans. Circuits Syst. Video Technol. **33**(8), 4096–4107 (2023) 2, 4

56. Zhang, P., Wang, Y., Liu, Y., Tu, Z., Lu, H.: Magic tokens: Select diverse tokens for multi-modal object re-identification. In: CVPR. pp. 17117–17126 (2024) 2

57. Zhang, Q., Lai, C., Liu, J., Huang, N., Han, J.: Fmcnet: Feature-level modality compensation for visible-infrared person re-identification. In: CVPR. pp. 7339–7348 (2022) 2

58. Zhang, Y., Wang, H.: Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification. In: CVPR. pp. 2153–2162 (2023) 10

59. Zhang, Y., Yan, Y., Lu, Y., Wang, H.: Towards a unified middle modality learning for visible-infrared person re-identification. In: ACM MM. pp. 788–796 (2021) 4

60. Zhang, Z., Xie, Y., Li, D., Zhang, W., Tian, Q.: Learning to align via wasserstein for person re-identification. IEEE Transactions on Image Processing **29**, 7104–7116 (2020) 2

61. Zou, C., Chen, Z., Cui, Z., Liu, Y., Zhang, C.: Discrepant and multi-instance proxies for unsupervised person re-identification. In: ICCV. pp. 11058–11068 (2023) 4

62. Zuo, J., Zhou, H., Nie, Y., Zhang, F., Guo, T., Sang, N., Wang, Y., Gao, C.: Ufinebench: Towards text-based person retrieval with ultra-fine granularity. In: CVPR. pp. 22010–22019 (2024) 4