International Journal of Computer Vision (2025) 133:3129–3148
https://doi.org/10.1007/s11263-024-02322-1

Exploring Homogeneous and Heterogeneous Consistent Label Associations for Unsupervised Visible-Infrared Person ReID

Lingfeng He[1] · De Cheng[1] · Nannan Wang[1] · Xinbo Gao[2]

Received: 25 April 2024 / Accepted: 29 November 2024 / Published online: 26 December 2024

Abstract

Unsupervised visible-infrared person re-identification (USL-VI-ReID) endeavors to retrieve pedestrian images of the same identity from different modalities without annotations. While prior work focuses on establishing cross-modality pseudo-label associations to bridge the modality-gap, they ignore maintaining the instance-level homogeneous and heterogeneous consistency between the feature space and the pseudo-label space, resulting in coarse associations. In response, we introduce a Modality-Unified Label Transfer (MULT) module that simultaneously accounts for both homogeneous and heterogeneous fine-grained instance-level structures, yielding high-quality cross-modality label associations. It models both homogeneous and heterogeneous affinities, leveraging them to quantify the inconsistency between the pseudo-label space and the feature space, subsequently minimizing it. The proposed MULT ensures that the generated pseudo-labels maintain alignment across modalities while upholding structural consistency within intra-modality. Additionally, a straightforward plug-and-play Online Cross-memory Label Refinement (OCLR) module is proposed to further mitigate the side effects of noisy pseudo-labels while simultaneously aligning different modalities, coupled with an Alternative Modality-Invariant Representation Learning (AMIRL) framework. Experiments demonstrate that our proposed method outperforms existing state-of-the-art USL-VI-ReID methods, highlighting the superiority of our MULT in comparison to other cross-modality association methods. Code is available at https://github.com/FranklinLingfeng/code_for_MULT.

Keywords Homogeneous and heterogeneous consistency · Label associations · Unsupervised visible-infrared person re-identification

1 Introduction

Visible-infrared person re-identification (VI-ReID) Lin et al. (2022); Zhang et al. (2021b); Wu et al. (2017); Ye et al. (2021b, 2020); Wu and Ye (2023) aims at retrieving the same person from a set of visible/infrared gallery images when given an image from another modality. It has garnered growing interest due to its practical applications in intelligent surveillance systems. Existing VI-ReID methods have achieved remarkable performance with deep neural networks Pu et al. (2020); Jia et al. (2020); Hao et al. (2019); Mao et al. (2017); Li et al. (2022b); Liu et al. (2022). However, these works are based on datasets with modality-shared annotations, which are labor-intensive and time-consuming to obtain in real-world scenarios. To relieve such issues, we investigate the unsupervised solution for VI-ReID.

Although there are several unsupervised single-modality ReID methods that have achieved excellent performance (*e.g.*, Cluster-Contrast Dai et al. (2023), ISE Zhang et al. (2022b), PPLR Cho et al. (2022)), the direct application of these methods to VI-ReID scenarios presents a formidable challenge due to the substantial modality-gap. Images of the same pedestrian from different modalities cannot be assigned the same pseudo-label using conventional clustering-based methods. Therefore, associating

Communicated by Yasushi Yagi.

✉ De Cheng
dcheng@xidian.edu.cn

✉ Nannan Wang
nnwang@xidian.edu.cn

Lingfeng He
lfhe@stu.xidian.edu.cn

Xinbo Gao
gaoxb@cqupt.edu.cn

[1] Xidian University, Xi'an 710071, China

[2] Chongqing University of Posts and Telecommunications, Chongqing 400065, China

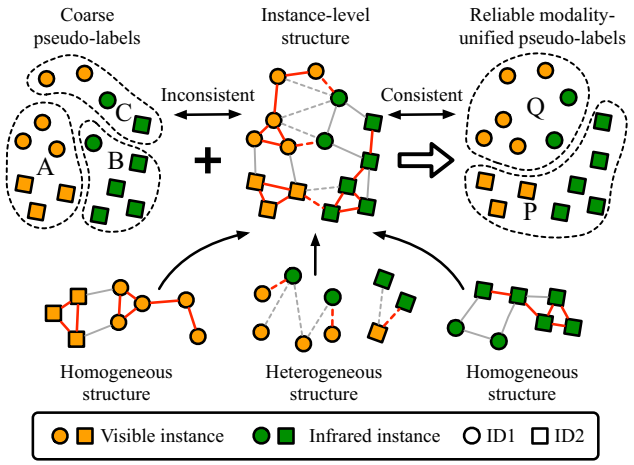

Fig. 1 Illustration of our idea. Different colors denote different modalities, and different shapes denote different identities. The red lines represent **higher affinities** and the gray lines represent **lower affinities**. Our Modality-Unified Label Transfer takes into account instance-level structures to establish homogeneous and heterogeneous structurally consistent label associations and generate reliable modality-unified pseudo-labels for network training

cross-modality pseudo-labels is necessary for unsupervised VI-ReID. Several attempts Wang et al. (2022); Cheng et al. (2023b, a); Wu and Ye (2023) have been made to associate the cross-modality pseudo-labels. PGM Wu and Ye (2023) and MBCCM Cheng et al. (2023a) adopt graph matching to interlink cross-modality clusters from the global perspective. However, they overlook the complicated, fine-grained structural information at the instance level, consequently resulting in coarse associations. To utilize the instance-level information, OTLA Wang et al. (2022) formulates the label assignment between instance and cross-modality clusters as an Optimal Transport (OT) problem. Nevertheless, it neglects the homogeneous structural consistency, leading to a large amount of intra-cluster instances in one modality being dispersed across multiple clusters in another modality. Based on the above analysis, we utilize instance-level pairwise relationships to establish reliable cross-modality label associations that maintain both homogeneous and heterogeneous structural consistency (as shown in Fig. 1).

Specifically, we propose a Modality-Unified Label Transfer (MULT) module (Fig. 2a), which exploits the full potential of both homogeneous and heterogeneous instance-level structures to associate cross-modality pseudo-labels. To begin, our MULT excavates homogeneous and heterogeneous structural information by modeling affinities derived from pairwise instance relationships in feature space. These affinities are then utilized to define both homogeneous and heterogeneous inconsistency between the pseudo-label space and the feature space from a global perspective. Subsequently, MULT transfers the pseudo-labels guided by the calculated affinities, with the primary aim of minimizing the inconsistency terms. During the label transfer process, each instance communicates its pseudo-label information with both its intra-modality and cross-modality counterparts. Such transfer strategy leverages detailed instance-level relationships, facilitating more precise associations compared to the direct associations of clusters. Simultaneously, the pseudo-labels preserve the homogeneous structure in feature space by minimizing the homogeneous inconsistency terms. Furthermore, the transferred soft pseudo-labels involve information between instances and multiple intra-modality and cross-modality clusters and thus are more suitable for mining discriminative features compared to hard labels. Extensive experiments indicate that our MULT provides more suitable supervision signals for training compared to other cross-association methods.

To achieve modality alignment based on pseudo-labels derived from MULT, we introduce an Online Cross-memory Label Refinement (OCLR) module (Fig. 2c), complemented by an Alternative Modality-Invariant Representation Learning (AMIRL) framework (Fig.2b). Specifically, our OCLR module is a straightforward yet effective plug-and-play component designed to alleviate the impact of the inevitable noisy pseudo-labels while further reducing the modality-gap. Specifically, it learns self-consistency among the predictions from multi-memory prototypes. Our AMIRL framework conducts contrastive learning based on both intra-modality and cross-modality memory banks. Two auxiliary memory banks are constructed to collaboratively learn the intra-modality structure with the initial two intra-modality memory banks, which play a role in mutually correcting each other. Furthermore, an alternative training scheme is proposed to multigate the influence of the inconsistency between the pseudo-labels from MULT in different directions. Experimental results highlight the effectiveness of our OCLR, showcasing its applicability in various cross-modality label association methods. Our main contributions can be summarized as follows:

- We propose a MULT module that considers instance-level context structures to provide homogeneous and heterogeneous consistent cross-modality pseudo-label associations for network training. The generated pseudo-labels exhibit cross-modality alignment while containing rich intra-modality information.
- We design a straightforward plug-and-play OCLR module for learning cross-memory self-consistency online, coupled with an AMIRL framework that fully exploits the supervision signals from MULT. The OCLR module effectively alleviates the side-effects of the noisy labels while mitigating the modality-gap.
- Extensive experimental results demonstrate the effectiveness of the proposed method, showing higher-quality

label associations across-modality instances, and superior recognition performances to state-of-the-art USL-VI-ReID methods on SYSU-MM01 and RegDB datasets.

2 Related Work

2.1 Supervised VI-ReID

Supervised VI-ReID pays attention to retrieving pedestrian images from both visible and infrared cameras. Existing methods mainly focus on aligning two modalities at the image-level and feature-level. Some methods Mao et al. (2017); Dai et al. (2018); Choi et al. (2020) bridge the modality-gap by transferring images from one modality to the other based on Generative Adversarial Networks (GANs). However, these methods face the issue of high computational cost and the inevitable noise in generated images, thus degenerating the model performance. CA Ye et al. (2021a) proposes a simple random channel argumentation for visible images to bridge the gap between cross-modality images. For aligning modalities at feature-level, some methods design novel network structures based on two-stream CNN with deep metric learning Liu et al. (2022); Wu et al. (2021); Ye et al. (2020); Zhang and Wang (2023); Wu et al. (2023) to excavate modality-invariant features. DDAG Ye et al. (2020) mines both intra-modality part-level and cross-modality graph-level contextual information through a dynamic dual-attention aggregation learning method. MPANet Wu et al. (2021) proposes a modality alleviation module and a pattern alignment module to extract part-level discriminative features. PartMix Kim et al. (2023) synthesizes positive and negative by mixing the part descriptors across modalities to enhance part-level feature learning. However, the aforementioned methods rely heavily on human-annotated cross-modality associations, which require expensive labor costs and are not always available in real scenarios. Thus we focus on learning modality-invariant features without human annotations.

2.2 Unsupervised Single-Modality ReID

Unsupervised single-modality ReID aims to learn discriminative representations for unannotated person images. Recent unsupervised methods Lin et al. (2020); Wang and Zhang (2020); Lin et al. (2019); Yu et al. (2019); Wang et al. (2021); Li et al. (2022); Zhang et al. (2023); Dai et al. (2023); Zhang et al. (2022b) follow a self-training pipeline that alternates between generating pseudo-labels and training the network. The mainstream methods for unsupervised ReID include pseudo-label refinement-based methods and memory bank-based methods. To alleviate the impact of noisy labels, MMT Ge et al. (2020a) utilizes the Mean-Teacher Tarvainen and Valpola (2017) model to revise pseudo-labels online. RLCC Zhang et al. (2021a) propagates pseudo-labels between different iterations through cluster consensus and generates temporally consistent pseudo-labels. PPLR Cho et al. (2022) proposes refining the pseudo-labels of global features by ensembling the predictions of part features based on the global-part agreement score. Memory bank-based methods establish memories capable of managing multi-prototypes and facilitate contrastive learning during training. SPCL Ge et al. (2020) constructs a hybrid memory that maintains both instance-level and cluster-level prototypes with a self-paced learning strategy. Cluster-Contrast Dai et al. (2023) and ISE Zhang et al. (2022b) store a unique prototype for each cluster to preserve the updating consistency. DCMIP Zhou et al. (2023) manages discrepant cluster memories and a multi-instance memory to excavate multifaceted information within clusters.

Following the memory-based methods, we construct intra-modality and cross-modality memory banks to perform contrastive learning for heterogeneous features.

2.3 Unsupervised VI-ReID

Unsupervised VI-ReID aims to learn modality-invariant and identity-discriminative features for cross-modality images without annotations. Existing methods Liang et al. (2021); Yang et al. (2022); Cheng et al. (2023a, b); Wu and Ye (2023) mainly focus on associating cross-modality pseudo-labels. H2H Liang et al. (2021) proposes an ISML loss aimed at enhancing alignment between reliable cross-modality instances. ADCA Yang et al. (2022) identifies highly associated cross-modality features and aggregates their corresponding memory prototypes according to instance-level pairwise similarities. To avoid biased label associations, OTLA Wang et al. (2022) establishes an optimal transport problem to assign cross-modality pseudo-labels uniformly for instances. To model the cluster-level cross-modality relationships from a global perspective, PGM Wu and Ye (2023) and MBCCM Cheng et al. (2023a) construct weighted bipartite graphs to associate cross-modality clusters and assign shared pseudo-labels for heterogeneous instances. CCLNet Chen et al. (2023) leverages the powerful semantic information from CLIP to provide richer supervision signals by incorporating a prompt learning stage. GRU Yang et al. (2023) proposes a CAE module to embed the information of hierarchical domain memories while achieving remarkable performance.

In contrast to the above-mentioned methods, our MULT simultaneously utilizes the homogeneous and heterogeneous instance-level structures to provide reliable cross-modality label associations.

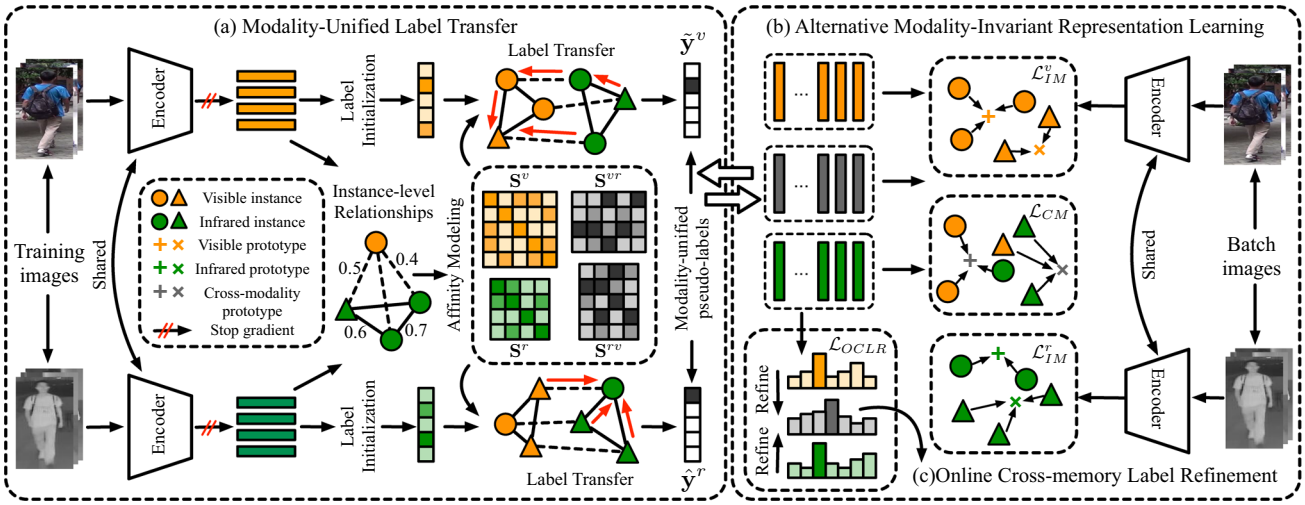

Fig. 2 Framework of our proposed method. Different colors indicate different modalities. Our method alternates pseudo-label generation (Modality-Unified Label Transfer (MULT **a**, described in Sec.3.1)) and network training (including Alternative Modality-Invariant Representation Learning (AMIRL **b**, described in Sec.3.2) and Online Cross-memory Label Refinement (OCLR **c**, described in Sec.3.3)). MULT provides homogeneous and heterogeneous consistent pseudo-labels as supervision signals. During training, AMIRL leverages memory banks to perform contrastive learning with an alternative scheme and OCLR utilizes predictions from different memories to alleviate the effect of the noisy labels

2.4 Affinity-Based Person ReID

In the person ReID task, some works Chen et al. (2018); Shen et al. (2018b, a); Luo et al. (2019); Lu et al. (2020); Li et al. (2022b); Ye et al. (2020) pay attention to the detailed relationships between pairwise instances. SSFT Luo et al. (2019) proposes a feature transfer module to facilitate the optimization of group-wise similarities for single-modality ReID. cm-SSFT Lu et al. (2020) models both intra-modality and cross-modality affinities to generate modality-shared features from modality-specific features. CIFT Li et al. (2022b) proposes a novel graph module to simulate the unbalanced modality distribution.

Inspired by the widely-used GCN Kipf and Welling (2016) and the above affinity-based methods, we model affinities in feature space to fully exploit instance-level relationships. Then we aim to obtain pseudo-labels that preserve instance-level contextual information.

3 Methodology

Given a visible-infrared dataset $\mathcal{X} = \{\mathcal{X}^v, \mathcal{X}^r\}$, $\mathcal{X}^v = \left\{\mathbf{x}_i^v | i = 1, 2, \cdots, N^v\right\}$ and $\mathcal{X}^r = \left\{\mathbf{x}_i^r | i = 1, 2, \cdots, N^r\right\}$ represents the visible and infrared datasets with N^v and N^r images, respectively. In the context of the USL-VI-ReID task, our objective is to train a deep neural network $f_\theta(\cdot)$ to project an image $\mathbf{x}_i$ from the dataset $\mathcal{X}$ into an embedding space $\mathcal{F}_\theta$ and derive a d-dimensional modality-invariant representation $\mathbf{f}_i = f_\theta(\mathbf{x}_i) \in \mathbb{R}^d$.

We employ the two stream encoder $f_\theta(\cdot)$ (*e.g.*, ResNet50 He et al. (2016)) as backbone to extract visible features $\left\{\mathbf{f}_i^v | i = 1, 2, \cdots, N^v\right\}$ and infrared features $\left\{\mathbf{f}_i^r | i = 1, 2, \cdots, N^r\right\}$. In the initial training stage, following Yang et al. (2022); Cheng et al. (2023a); Wu and Ye (2023), we employ the Dual-Contrastive Learning (DCL) framework Yang et al. (2022) to facilitate intra-modality contrastive learning as our baseline.

Our proposed method is employed during the second training stage. The framework is illustrated in Fig. 2. Following the well-developed unsupervised methods Ge et al. (2020); Dai et al. (2023); Zhang et al. (2022b); Yang et al. (2022), we alternate between pseudo-label generation and network training. During the pseudo-label generation stage, following Yang et al. (2022); Wu and Ye (2023); Cheng et al. (2023a), we first utilize DBSCAN Ester et al. (1996) to cluster features. Two intra-modality memory banks $\tilde{\mathbf{M}}^v \in \mathbb{R}^{K^v \times d}$ and $\tilde{\mathbf{M}}^r \in \mathbb{R}^{K^r \times d}$ are initialized by the cluster centroids of their corresponding modalities, where K^v and K^r denotes the number of clusters in the visible modality and the infrared modality, respectively. Then the proposed Modality-Unified Label Transfer (MULT, Fig. 2a) establishes reliable cross-modality label associations and generates soft modality-unified pseudo-labels for network training. During the training stage, we propose an Alternative Modality-Invariant Representation learning (AMIRL, Fig. 2b) framework to fully leverage the supervisory signals from MULT through contrastive learning, coupled with an alternative scheme. Additionally, an Online Cross-memory Label refinement (OCLR, Fig. 2c) module is proposed to further alleviate the influence of noisy labels while reducing the

modality discrepancy. The entire framework includes two training modes, *i.e.*, V-based and R-based, where V-based denotes we utilize pseudo-labels in the visible label space to guide network training, and vice versa.

3.1 Modality-Unified Label Transfer

Motivated by the affinity-based ReID methods Luo et al. (2019); Lu et al. (2020); Li et al. (2022b), Our MULT models instance-wise affinities to generate structurally consistent pseudo-labels. In MULT, every instance holds two types of pseudo-labels, *i.e.*, the intra-modality label (labels from the label space corresponding to the instance's modality) and the cross-modality label (labels from the label space corresponding to another modality). Formally, We use $\mathbf{y}^e = \{\tilde{\mathbf{y}}^e, \hat{\mathbf{y}}^e\}$ to denote the pseudo-labels, where $\tilde{\mathbf{y}}^e$ denotes the intra-modality labels and $\hat{\mathbf{y}}^e$ denotes the cross-modality labels. $e = \{v, r\}$ indicates visible and infrared modality, respectively, and $\mathbf{y}_i^e$ denotes the pseudo-label of i-th instance in modality e. (*e.g.*, $\tilde{\mathbf{y}}^v$ denotes pseudo-labels for visible instances in visible label space and $\hat{\mathbf{y}}^v$ denotes pseudo-labels for visible instances in infrared label space).

Our MULT includes two directions, *i.e.*, V2R and R2V. The V2R MULT involves the transfer between visible intra-modality labels $\tilde{\mathbf{y}}^v$ and infrared cross-modality labels $\hat{\mathbf{y}}^r$ within the visible label space, and verse vice. For convenience, we only describe the V2R MULT in detail.

Affinity Modeling. To incorporate instance-level relationships, we model the homogeneous and heterogeneous affinities, denoted as $\mathbf{S}^{ho(e)}$ and $\mathbf{S}^{he}$. $\mathbf{S}^{ho(e)} = \{\mathbf{S}^{ho(v)}, \mathbf{S}^{ho(r)}\}$ denotes the affinities within visible and infrared modalities, respectively. To enhance the consistency between the homogeneous affinities and the clustering results, we employ the *Jaccard Similarity* Zhong et al. (2017) for clustering to model ***the homogeneous affinities***:

$$\mathbf{S}_{ij}^{ho(v)} = \frac{|\mathcal{R}(\mathbf{f}_i^v, \kappa) \cap \mathcal{R}(\mathbf{f}_j^v, \kappa)|}{|\mathcal{R}(\mathbf{f}_i^v, \kappa) \cup \mathcal{R}(\mathbf{f}_j^v, \kappa)|}, \mathbf{S}^{ho(v)} \in \mathbb{R}^{N^v \times N^v}, \tag{1}$$

$$\mathbf{S}_{ij}^{ho(r)} = \frac{|\mathcal{R}(\mathbf{f}_i^r, \kappa) \cap \mathcal{R}(\mathbf{f}_j^r, \kappa)|}{|\mathcal{R}(\mathbf{f}_i^r, \kappa) \cup \mathcal{R}(\mathbf{f}_j^r, \kappa)|}, \mathbf{S}^{ho(r)} \in \mathbb{R}^{N^r \times N^r}, \tag{2}$$

where $\mathcal{R}(\mathbf{f}_i, \kappa)$ is the κ-reciprocal nearest neighbors Zhong et al. (2017) of $\mathbf{f}_i$, *i.e.*, $\mathcal{R}(\mathbf{f}_i, \kappa) = \{\mathbf{g_i} | (\mathbf{g_i} \in \mathbb{N}(\mathbf{f}_i, \kappa)) \wedge (\mathbf{f}_i \in \mathbb{N}(\mathbf{g_i}, \kappa))\}$, and $\mathbb{N}(\mathbf{f}_i, \kappa))$ denotes the κ-nearest neighbors of the probe $\mathbf{f}_i$, $|\cdot|$ is the cardinality of a set. $\mathbf{S}_{ij}^{ho}$ can be regarded as the affinity between homogeneous features $\mathbf{f}_i$ and $\mathbf{f}_j$.

Due to the substantial modality discrepancy, it is inappropriate to directly model the heterogeneous affinity by the *Jaccard Similarities* between cross-modality features as Eq.1 and Eq.2. It should take into account both the cross-modality instance-level relationships and the misalignment between distributions of the two modalities. Thus affinity modeling can be conceptualized as a transition process between instances from two different distributions, which can be formulated as an optimal transport problem. Each instance in the dataset should be treated equally, meaning the marginal distributions of instances from both modalities follow uniform distributions. Thus we model ***the heterogeneous affinites*** by solving the Optimal Transport (OT) plan according to the transport cost between cross-modality features, which can be formulated as:

$$\begin{aligned} &\min_{\mathbf{S}^{he}} \langle \mathbf{S}^{he}, \mathbf{C}^{he} \rangle + \frac{1}{\lambda} \langle \mathbf{S}^{he}, -\log(\mathbf{S}^{he}) \rangle. \\ &s.t. \quad \mathbf{S}^{he} \mathbb{1} = \mathbb{1} \cdot \frac{1}{N^v}, \mathbf{S}^{he\top} \mathbb{1} = \mathbb{1} \cdot \frac{1}{N^r}, \end{aligned} \tag{3}$$

where $\langle \cdot \rangle$ denotes the Frobenius dot-product, and $\mathbb{1}$ is an all in 1 vector. $\mathbf{S}^{he} \in \mathbb{R}^{N^v \times N^r}$ denotes the transport plan. $\mathbf{C}^{he} \in \mathbb{R}^{N^v \times N^r}$ is the cost matrix constructed by computing the Euclidean distance of heterogeneous features, *i.e.*, $\mathbf{C}_{ij}^{he} = \|\mathbf{f}_i^v - \mathbf{f}_j^r\|_2^2$. The first term in the objective function in Eq.3 indicates the total transport cost and the second term is the Entropic Regularization Courty et al. (2016). The constraints ensure that N^v visible instances are uniformly distributed to N_r infrared instances in the transport plan, and verse vice. Actually, few instances in one modality inherently exhibit closer proximity to another modality compared to other instances, and such an issue is avoided by the above constraints. The optimal solution $\mathbf{S}^{he*}$ of Eq.3 can be obtained by the Sinkhorn-Knopp algorithm Cuturi (2013), and $\mathbf{S}_{ij}^{he*}$ can be regarded as the affinity between heterogeneous features $\mathbf{f}_i^v$ and $\mathbf{f}_j^r$.

We then derive the visible-to-infrared affinity matrix $\mathbf{S}^{he(vr)} = \mathbf{S}^{he*} \in \mathbb{R}^{N^v \times N^r}$ and the infrared-to-visible affinity matrix $\mathbf{S}^{he(rv)} = \mathbf{S}^{he*\top} \in \mathbb{R}^{N^r \times N^v}$ from the heterogeneous affinities $\mathbf{S}^{he*}$. We perform row normalization on the above affinities and obtain the final affinity matrices $\mathbf{S} = \{\mathbf{S}^{ho(v)}, \mathbf{S}^{ho(r)}, \mathbf{S}^{he(vr)}, \mathbf{S}^{he(rv)}\}$, which are subsequently employed to define the structural inconsistency.

Inconsistency Formulation. We leverage the above affinities to define the inconsistency terms for pseudo-labels. A minor degree of structural inconsistency implies that the higher the affinity between two instances, the smaller the distance between their pseudo-labels. Consequently, we define the product of the Euclidean distance between pseudo-labels of pairwise instances and their affinity as their pairwise structural inconsistency. From a global perspective, for all instances in both modalities, ***the homogenous inconsistency*** can be formulated as follows:

$$\mathcal{I}_{ho}^{v}(\tilde{\mathbf{y}}^{v}) = \sum_{i=1}^{N^{v}}\sum_{j=1}^{N^{v}} \mathbf{S}_{ij}^{ho(v)} \|\tilde{\mathbf{y}}_{i}^{v} - \tilde{\mathbf{y}}_{j}^{v}\|_{2}^{2}, \tag{4}$$

$$\mathcal{I}_{ho}^{r}(\hat{\mathbf{y}}^{r}) = \sum_{i=1}^{N^{r}}\sum_{j=1}^{N^{r}} \mathbf{S}_{ij}^{ho(r)} \|\hat{\mathbf{y}}_{i}^{r} - \hat{\mathbf{y}}_{j}^{r}\|_{2}^{2}. \tag{5}$$

The above inconsistency term can be regarded as a weighted sum of pairwise distances between the soft pseudo-labels. The higher the affinities between two instances in feature space, the more their pseudo-label distance contributes to the computation of the inconsistency. ***The heterogeneous inconsistency*** can be formulated in a similar manner:

$$\mathcal{I}_{he}^{v}(\tilde{\mathbf{y}}^{v}) = \sum_{i=1}^{N^{v}}\sum_{j=1}^{N^{r}} \mathbf{S}_{ij}^{he(vr)} \|\tilde{\mathbf{y}}_{i}^{v} - \hat{\mathbf{y}}_{j}^{r}\|_{2}^{2}, \tag{6}$$

Nonetheless, solely minimizing the above two inconsistency terms can lead to a collapse, where all instances end up with nearly identical labels. This is because these two inconsistency statements do not consider the fact that pairwise instances with low affinities should have distinct labels. Such a reduction in label diversity can negatively impact network training. To avoid such collapse, a label initialization approach that ensures the diversity of the pseudo-labels is necessary.

Specifically, the intra-modality labels are initialized according to the cluster centroids, and the cross-modality labels are initialized by the widely-used OTLA Wang et al. (2022) method. To fully utilize the relationships between instances and each cluster centroid, we use the soft probability distribution from the corresponding memory bank as initial visible intra-modality labels $\tilde{\mathbf{y}}^{v}$:

$$\tilde{\mathbf{y}}^{v} \in \mathbb{R}^{N^{v}\times K^{v}}, \tilde{\mathbf{y}}_{i}^{v}(0) = \mathbf{P}(\mathbf{f}_{i}^{v}|\tilde{\mathbf{M}}^{v}, \tau) \in \mathbb{R}^{K^{v}}, \tag{7}$$

where $\mathbf{P}(\mathbf{f}|\mathbf{M}, \tau)$ denotes the probability distribution output from memory bank $\mathbf{M}$ for feature $\mathbf{f}$ with the temperature factor τ, which can be formulated as:

$$P_{j}(\mathbf{f}|\mathbf{M}, \tau) = \frac{\exp\left(\mathbf{f}^{\top}\mathbf{c}_{j}/\tau\right)}{\sum_{k=1}^{K}\exp\left(\mathbf{f}^{\top}\mathbf{c}_{k}/\tau\right)}, \tag{8}$$

$$\mathbf{P}(\mathbf{f}|\mathbf{M}, \tau) = [P_{1}(\mathbf{f}|\mathbf{M}, \tau), \cdots, P_{K}(\mathbf{f}|\mathbf{M}, \tau)], \tag{9}$$

where K denotes the number of prototypes in the memory $\mathbf{M}$, $\mathbf{c}_{j}$ and $\mathbf{c}_{k}$ represents j-th and k-th cluster prototypes in $\mathbf{M}$. Following Wang et al. (2022); Cheng et al. (2023b), we utilize Optimal Transport Label Assignment (OTLA) to initialize infrared cross-modality labels $\hat{\mathbf{y}}^{r}$:

$$\min_{\mathbf{P}} \langle \mathbf{P}, \mathbf{C}\rangle + \frac{1}{\lambda}\langle \mathbf{P}, -\log(\mathbf{P})\rangle.$$
$$s.t. \quad \mathbf{P}\mathbb{1} = \mathbb{1}\cdot\frac{1}{N^{r}}, \quad \mathbf{P}^{\top}\mathbb{1} = \mathbb{1}\cdot\frac{1}{K^{v}}, \tag{10}$$

where N^{r} denotes the number of infrared instances, and K^{v} denotes the number of visible clusters. $\mathbf{C} \in \mathbb{R}^{N^{r}\times K^{v}}$ represents the cost matrix and $\mathbf{P} \in \mathbb{R}^{N^{r}\times K^{v}}$ represents the transport plan. We utilize the Euclidean distance between infrared instances and visible prototypes from $\tilde{\mathbf{M}}^{v}$ to calculate the cost, where $\mathbf{C}_{ij} = \|\mathbf{f}_{i}^{r} - \tilde{\mathbf{m}}_{j}^{v}\|_{2}^{2}$. $\mathbf{f}_{i}^{r}$ is the i-th feature vector in training set and $\tilde{\mathbf{m}}_{j}^{v}$ is the j-th prototype in memory $\tilde{\mathbf{M}}^{v}$. The initialized $\hat{\mathbf{y}}^{r}(0) \in \mathbb{R}^{N^{r}\times K^{v}}$ is the one-hot encoding form of the optimal solution $\mathbf{P}^{*}$ of Eq.10, which is formulated as follows:

$$\hat{\mathbf{y}}_{ik}^{r}(0) = \begin{cases} 1, & \text{if } \ k = \arg\max_{j} \mathbf{P}_{ij}^{*} \\ 0, & \text{otherwise} \end{cases}, \quad \hat{\mathbf{y}}_{i}^{r}(0) \in \mathbb{R}^{K^{v}}. \tag{11}$$

In the optimization process, we aim to ensure that the transferred labels do not deviate too far from the initial labels, thereby guaranteeing diversity in the transferred labels. Therefore, we propose the ***self-inconsistency*** to constraint the Euclidean distance between the transferred labels and the initial coarse labels:

$$\mathcal{I}_{self}^{v}(\tilde{\mathbf{y}}^{v}) = \sum_{i=1}^{N^{v}} \|\tilde{\mathbf{y}}_{i}^{v} - \tilde{\mathbf{y}}_{i}^{v}(0)\|_{2}^{2}. \tag{12}$$

$$\mathcal{I}_{self}^{r}(\hat{\mathbf{y}}^{r}) = \sum_{i=1}^{N^{r}} \|\hat{\mathbf{y}}_{i}^{r} - \hat{\mathbf{y}}_{i}^{r}(0)\|_{2}^{2}. \tag{13}$$

We minimize the above three terms of inconsistency, aiming to derive the pseudo-labels $\tilde{\mathbf{y}}^{v}$ and $\hat{\mathbf{y}}^{r}$ with the lowest inconsistency. This optimization can be formulated as follows:

$$\min_{\tilde{\mathbf{y}}^{v}} \quad \mathcal{I}_{ho}^{e}(\tilde{\mathbf{y}}^{v}) + \alpha\mathcal{I}_{self}^{e}(\tilde{\mathbf{y}}^{v}) + (1-\alpha)\mathcal{I}_{he}^{e}(\tilde{\mathbf{y}}^{v}), \tag{14}$$

$$\min_{\hat{\mathbf{y}}^{r}} \quad \mathcal{I}_{ho}^{e}(\hat{\mathbf{y}}^{r}) + \alpha\mathcal{I}_{self}^{e}(\hat{\mathbf{y}}^{r}) + (1-\alpha)\mathcal{I}_{he}^{e}(\hat{\mathbf{y}}^{r}), \tag{15}$$

where α is a trade-off parameter.

Label Transfer. The above optimization problem can be solved by the LGC algorithm Zhou et al. (2003). Specifically, the visible and infrared modality-unified labels are updated alternately as Eq.16 and Eq.17 until convergence:

$$\begin{aligned}\tilde{\mathbf{y}}^v(t+1)^* &= (1-\alpha)\mathbf{S}^{he(vr)}\cdot\hat{\mathbf{y}}^r(t)+\alpha\tilde{\mathbf{y}}^v(0),\\ \tilde{\mathbf{y}}^v(t+1) &= \tfrac{1}{2}\mathbf{S}^{ho(v)}\cdot\tilde{\mathbf{y}}^v(t+1)^*+\tfrac{1}{2}\tilde{\mathbf{y}}^v(t+1)^*,\end{aligned} \tag{16}$$

$$\begin{aligned}\hat{\mathbf{y}}^r(t+1)^* &= (1-\alpha)\mathbf{S}^{he(rv)}\cdot\tilde{\mathbf{y}}^v(t)+\alpha\hat{\mathbf{y}}^r(0),\\ \hat{\mathbf{y}}^r(t+1) &= \tfrac{1}{2}\mathbf{S}^{ho(r)}\cdot\hat{\mathbf{y}}^r(t+1)^*+\tfrac{1}{2}\hat{\mathbf{y}}^r(t+1)^*,\end{aligned} \tag{17}$$

where $\mathbf{y}^e(t)=\{\tilde{\mathbf{y}}^v(t),\hat{\mathbf{y}}^r(t)\}$ denotes the labels in $t-th$ iteration during the optimization process. $\mathbf{y}^e(t+1)^*$ denotes the intermediate variables during the update process in one iteration. The upper terms in Eq.16 and Eq.17 can be regarded as a cross-modality transfer process, while the lower terms stand for a manner to further compensate the transferred pseudo-labels utilizing intra-modality structural information.

In detail, take the pseudo-labels $\tilde{\mathbf{y}}^v$ in visible modality, the inconsistency formulation Eq.14 can be expressed in the form of vectors and matrices:

$$\begin{aligned}\mathcal{I}(\tilde{\mathbf{y}}^v) &= \alpha\mathbf{tr}((\tilde{\mathbf{y}}^v-\tilde{\mathbf{y}}^v(0))^\top(\tilde{\mathbf{y}}^v-\tilde{\mathbf{y}}^v(0)))\\ &+(1-\alpha)\mathbf{tr}((\tilde{\mathbf{y}}^v-\mathbf{S}^{he(vr)}\cdot\hat{\mathbf{y}}^r)^\top(\tilde{\mathbf{y}}^v-\mathbf{S}^{he(vr)}\cdot\hat{\mathbf{y}}^r))\\ &+\mathbf{tr}(\tilde{\mathbf{y}}^{v\top}(\mathbf{I}_{N^v}-\mathbf{S}^{ho(v)})\tilde{\mathbf{y}}^v),\end{aligned} \tag{18}$$

where $\mathbf{I}_{N^v}$ denotes the $N^v\times N^v$ identity matrix and $\mathbf{tr}(\cdot)$ denotes the trace of matrix. To minimize the inconsistency $\mathcal{I}(\tilde{\mathbf{y}}^v)$ for visible pseudo-labels, we compute its partial derivative with respect to $\tilde{\mathbf{y}}^v$ and equate this partial derivative to zero, where $\frac{\partial\mathcal{I}(\tilde{\mathbf{y}}^v)}{\partial\tilde{\mathbf{y}}^v}=0$ Then $\tilde{\mathbf{y}}^v$ can be represented as:

$$\tilde{\mathbf{y}}^v=\frac{1}{2}\mathbf{S}^{ho(v)}\cdot\tilde{\mathbf{y}}^v+\frac{1}{2}\left[(1-\alpha)\mathbf{S}^{he(vr)}\cdot\hat{\mathbf{y}}^r+\alpha\tilde{\mathbf{y}}^v(0)\right]. \tag{19}$$

To enhance its convergence stability, we decompose Eq.19 into two steps, *i.e.*, cross-modality transfer, and intra-modality transfer, which is formulated as follows:

$$\begin{aligned}\tilde{\mathbf{y}}^v &\xleftarrow{\text{cross-modality transfer}} (1-\alpha)\mathbf{S}^{he(vr)}\cdot\hat{\mathbf{y}}^r+\alpha\tilde{\mathbf{y}}^v(0),\\ \tilde{\mathbf{y}}^v &\xleftarrow{\text{intra-modality transfer}} \frac{1}{2}\mathbf{S}^{ho(v)}\cdot\tilde{\mathbf{y}}^v+\frac{1}{2}\tilde{\mathbf{y}}^v.\end{aligned} \tag{20}$$

During each iterative optimization process, the pseudo-labels $\tilde{\mathbf{y}}^v$ are updated by alternating iterations of fixed points, which is formulated as Eq.16.

Nevertheless, the transferred pseudo-labels are excessively smoothed, which is detrimental to the network's ability to learn predictions with low entropy. Therefore, we derive the soft modality-unified labels by aggregating both hard labels and soft forms of the pseudo-labels, which can be formulated as:

Algorithm 1: V2R MULT

Input: $\mathbf{f}^v$: Visible features; $\mathbf{f}^r$: Infrared features;
Output:
Visible intra-modality labels $\tilde{\mathbf{y}}^v$;
Infrared cross-modality labels $\hat{\mathbf{y}}^r$;
Initialization:
Two intra-modality memory banks $\tilde{\mathbf{M}}^v$ and $\tilde{\mathbf{M}}^r$;
Intra-modality visible labels $\tilde{\mathbf{y}}^v(0)$ (Eq.7);
Cross-modality infrared labels $\hat{\mathbf{y}}^r(0)$ (Eq.10);
Affinity Modeling:
Intra-modality affinity matrices $\mathbf{S}^v$ and $\mathbf{S}^r$ (Eq.1 & Eq.2);
Cross-modality affinity matrices $\mathbf{S}^{vr}$ and $\mathbf{S}_{rv}$ (Eq.3);
Do Label Transfer (V2R):
$t=0;\epsilon_0=1e-2;\epsilon=1e6;$
while $\epsilon>\epsilon_0$ **do**
  Calculate $\tilde{\mathbf{y}}^v(t+1)$ according to Eq.16;
  Calculate $\hat{\mathbf{y}}^r(t+1)$ according to Eq.17;
  $\epsilon\leftarrow$ $\max(\|\tilde{\mathbf{y}}^v(t+1)-\tilde{\mathbf{y}}^v(t)\|_1,\|\hat{\mathbf{y}}^r(t+1)-\hat{\mathbf{y}}^r(t)\|_1)$;
  $t\leftarrow t+1$;
Update labels $\tilde{\mathbf{y}}^v\leftarrow\tilde{\mathbf{y}}^v(t)$; $\hat{\mathbf{y}}^r\leftarrow\hat{\mathbf{y}}^r(t)$;
Return: Soft labels $\tilde{\mathbf{y}}^v$ and $\hat{\mathbf{y}}^r$ (Eq.21).

$$\mathbf{y}^e=\beta\cdot\mathbf{y}^e_{hard}+(1-\beta)\cdot\mathbf{y}^e_{soft}, \tag{21}$$

where β is a parameter to control the smoothness of the pseudo-label. $\mathbf{y}^e_{hard}$ and $\mathbf{y}^e_{soft}$ denotes the hard one-hot encoding pseudo-labels and the soft pseudo-labels obtained from the transferred labels $\mathbf{y}^e=\{\tilde{\mathbf{y}}^v,\hat{\mathbf{y}}^r\}$, respectively. The one-hot form $\mathbf{y}^e_{hard}$ is directly derived from the soft label $\mathbf{y}^e_{soft}$.

It is noteworthy that, the labels $\tilde{\mathbf{y}}^v_i$ and $\hat{\mathbf{y}}^r_i$ have a same dimension of K^v. The detailed algorithm flow of our V2R MULT is in Alg.1. The R2V MULT, which provides $\tilde{\mathbf{y}}^r$ and $\hat{\mathbf{y}}^v$ in infrared label space, holds a symmetrical form with V2R MULT.

3.2 Alternative Modality-Invariant Representation Learning

Guided by the modality-unified soft labels from the MULT module, we lay emphasis on alleviating the cross-modality discrepancy while keeping intra-modality consistency. We concurrently execute intra-modality and cross-modality contrastive learning. Additionally, we design an alternative scheme that alternates between V-based and R-based training modes.

To be specific, for V-based training mode, we construct a cross-modality memory bank $\mathbf{M}^c\in\mathbb{R}^{K^v\times d}$ and an auxiliary memory bank $\hat{\mathbf{M}}^r\in\mathbb{R}^{K^v\times d}$, which are initialized in

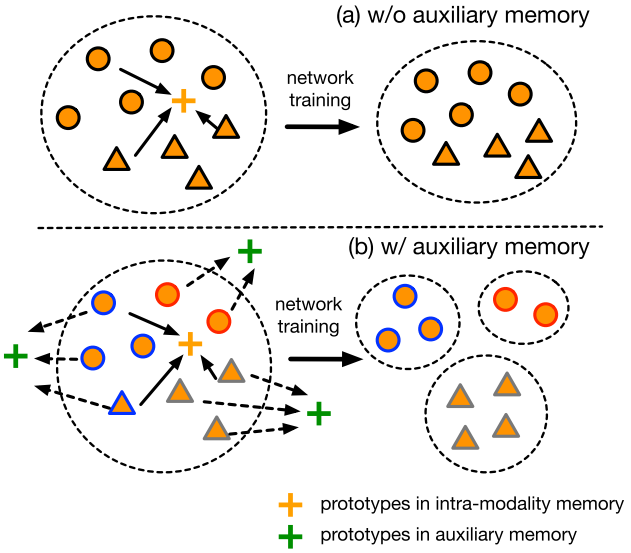

Fig. 3 Illustration of the role of the auxiliary memory. Different shapes denote different identities and different edge colors denote different cross-modality labels

the same manner as the visible intra-modality memory bank $\tilde{\mathbf{M}}^v$. $\mathbf{M}^c$ is updated by instances from both modalities and $\hat{\mathbf{M}}^r$ is updated by instances only from the infrared modality. The auxiliary memory bank $\hat{\mathbf{M}}^r$ aims to learn the infrared intra-modality structure depicted by its labels $\hat{\mathbf{y}}^r$ in visible label space, thereby serving as a compensatory mechanism for $\tilde{\mathbf{M}}^r$. In fact, instances from different identities are sometimes gathered in the same cluster with their intra-modality labels ($\tilde{\mathbf{y}}^r$), while they are distributed correctly into different clusters with their cross-modality labels ($\hat{\mathbf{y}}^r$). Thus the auxiliary memory, which reflects the structure of $\hat{\mathbf{y}}^r$, can assist the intra-modality learning. As shown in Flg.3, compared to utilizing prototypes solely from the intra-modality memory, leveraging prototypes from the auxiliary memory of another modality is advantageous for uncovering fine-grained structural details within clusters in both two modalities.

Following the widely-used memory-based methods Ge et al. (2020); Dai et al. (2023); Yang et al. (2022); Cheng et al. (2023), we carry out contrastive learning during forward-propagation (FP) and update the memory banks during backward-propagation (BP). The visible intra-modality contrastive loss can be formulated as:

$$\mathcal{L}_{IM}^{v} = \frac{1}{B}\sum_{i=1}^{B}\mathcal{L}_{ce}(\mathbf{P}(\mathbf{f}_i^v|\tilde{\mathbf{M}}^v,\tau),\tilde{\mathbf{y}}_i^v), \tag{22}$$

where $\mathcal{L}_{ce}$ denotes the standard cross-entropy loss and B denotes the input batch size. For infrared modality in V-based training, both $\tilde{\mathbf{y}}^r$ and $\hat{\mathbf{y}}^r$ are utilized jointly during contrastive learning with their corresponding memory $\tilde{\mathbf{M}}^r$ and $\hat{\mathbf{M}}^r$:

$$\begin{aligned}\mathcal{L}_{IM}^{r} = &\frac{1}{B}\sum_{i=1}^{B}\mathcal{L}_{ce}(\mathbf{P}(\mathbf{f}_i^r|\tilde{\mathbf{M}}^r,\tau),\tilde{\mathbf{y}}_i^r)\\ &+\frac{1}{B}\sum_{i=1}^{B}\mathcal{L}_{ce}(\mathbf{P}(\mathbf{f}_i^r|\hat{\mathbf{M}}^r,\tau),\hat{\mathbf{y}}_i^r).\end{aligned} \tag{23}$$

These two types of labels in two different label spaces play a mutual corrective role and facilitate intra-modality training. The total intra-modality contrastive loss can be formulated as follows:

$$\mathcal{L}_{IM} = \mathcal{L}_{IM}^{v} + \mathcal{L}_{IM}^{r}. \tag{24}$$

Besides, we perform cross-modality contrastive learning to constrain the relationships between instances in different modalities and their corresponding modality-agnostic cluster prototypes in $\mathbf{M}^c$ in the feature space:

$$\begin{aligned}\mathcal{L}_{CM} = &\frac{1}{B}\sum_{i=1}^{B}\mathcal{L}_{ce}(\mathbf{P}(\mathbf{f}_i^v|\mathbf{M}^c,\tau),\tilde{\mathbf{y}}_i^v)\\ &+\frac{1}{B}\sum_{i=1}^{B}\mathcal{L}_{ce}(P(\mathbf{f}_i^r|\mathbf{M}^c,\tau),\hat{\mathbf{y}}_i^r).\end{aligned} \tag{25}$$

During the BP stage, the memory banks are updated with the input features in a momentum strategy He et al. (2020); Dai et al. (2023):

$$\begin{aligned}
\tilde{\mathbf{m}}^v(\tilde{y}_i^v) &\leftarrow \mu\tilde{\mathbf{m}}^v(\tilde{y}_i^v)+(1-\mu)\mathbf{f}(\tilde{y}_i^v), \mathbf{f}\in\{\mathbf{f}_1^v,\cdots,\mathbf{f}_B^v\}\\
\tilde{\mathbf{m}}^r(\tilde{y}_i^r) &\leftarrow \mu\tilde{\mathbf{m}}^v(\tilde{y}_i^r)+(1-\mu)\mathbf{f}(\tilde{y}_i^r), \mathbf{f}\in\{\mathbf{f}_1^r,\cdots,\mathbf{f}_B^r\}\\
\mathbf{m}^c(\tilde{y}_i^v) &\leftarrow \mu\mathbf{m}^c(\tilde{y}_i^v)+(1-\mu)\mathbf{f}(\tilde{y}_i^v), \mathbf{f}\in\{\mathbf{f}_1^v,\cdots,\mathbf{f}_B^v\}\\
\mathbf{m}^c(\hat{y}_i^r) &\leftarrow \mu\mathbf{m}^c(\hat{y}_i^r)+(1-\mu)\mathbf{f}(\hat{y}_i^r), \mathbf{f}\in\{\mathbf{f}_1^r,\cdots,\mathbf{f}_B^r\}\\
\hat{\mathbf{m}}^r(\hat{y}_i^r) &\leftarrow \mu\hat{\mathbf{m}}^r(\hat{y}_i^r)+(1-\mu)\mathbf{f}(\hat{y}_i^r), \mathbf{f}\in\{\mathbf{f}_1^r,\cdots,\mathbf{f}_B^r\},
\end{aligned} \tag{26}$$

where μ is the momentum updating factor. $\mathbf{m}(y)$ is the $y-th$ prototype in the memory bank $\mathbf{M}$, and $\mathbf{f}(y)$ is the input feature with label y in current mini-batch. Note that $y=\{\tilde{y}_i^v,\tilde{y}_i^r,\hat{y}_i^r\}$ for memory update is the hard label form of the soft labels $\mathbf{y}=\{\tilde{\mathbf{y}}_i^v,\tilde{\mathbf{y}}_i^r,\hat{\mathbf{y}}_i^r\}$.

For R-based training, a cross-modality memory bank $\mathbf{M}^c\in\mathbb{R}^{K^r\times d}$ is constructed according to labels $\tilde{\mathbf{y}}^r$ and $\hat{\mathbf{y}}^v$ from R2V MULT. Meanwhile an auxiliary memory bank $\hat{\mathbf{M}}^{\mathbf{v}}\in\mathbb{R}^{K^r\times d}$ is initialized in the same manner as $\tilde{\mathbf{M}}^r$. It is updated by visible features based on $\hat{\mathbf{y}}^v$ in the infrared label space. The loss functions for R-based training can be written in a symmetric form in comparison to the V-based training.

The entire training process is described in Alg.2. We conduct an alternate scheme for the V-based training and the R-based training as the epoch goes on: when epoch-id

Algorithm 2: Training process in one epoch.

Input: epoch E; network f_θ; iteration number from T_0 to T_t in one epoch; batch size B;
1: Extract visible features $\{\mathbf{f}_i^v | i = 1, 2, \cdots, N^v\}$ and infrared features $\{\mathbf{f}_i^r | i = 1, 2, \cdots, N^r\}$;
2: Clustering features and initialize two intra-modality memory banks $\tilde{\mathbf{M}}^v$ and $\tilde{\mathbf{M}}^r$ based on the cluster centroids;
3: Do V2R MULT / R2V MULT according to Alg.1;
if $E\%2 == 0$: **then**
  1: Initialize cross-modality memory $\mathbf{M}^c$ and auxiliary memory $\hat{\mathbf{M}}^r$ according to the results from V2R MULT;
  2: **for** $t = T_0$ *to* T_t **do**
    Do V-based training;
else if $E\%2 == 1$ **then**
  1: Initialize cross-modality memory $\mathbf{M}^c$ and auxiliary memory $\hat{\mathbf{M}}^v$ according to the results from R2V MULT;
  2: **for** $t = T_0$ *to* T_t **do**
    Do R-based training;
4: **Output:** updated network f_θ.

$E\%2 == 0$, the V-based training is conducted; otherwise, the R-based training is conducted. In one epoch, the model is supervised by labels from MULT of a specific direction. In adjacent epochs, the model is supervised by labels from MULT of different directions. Such an alternative scheme mitigates the inconsistency between pseudo-labels from MULT of different directions (V2R / R2V).

3.3 Online Cross-memory Label Refinement

While the MULT module enables training under modality-unified supervision, it is essential to consider that due to the inherent noise in clustering and MULT, prolonged training with fixed supervised signals may cause the network to overfit incorrect labels.

To overcome this issue, we utilize the predictions from the intra-modality memory banks to refine the predictions from the cross-modality memory bank. Several reasons support the adoption of such refinement strategies: (1) The intra-modality memory banks are updated by features within their respective modalities, this update occurs at a slower rate compared to the cross-modality memory bank, resulting in more stable predictions. (2) The intra-modality and cross-modality memory banks are expected to output consistent predictions. In the visible modality, we adopt mutual refinement between the extracted features $\mathbf{f}_i^v$ from visible images and $\mathbf{f}_i^a$ from images with CA augmentation Ye et al. (2021a) to enhance constraints on their self-consistency. CA is a widely-used data augmentation strategy in VI-ReID Yang et al. (2022); Cheng et al. (2023a); Wu and Ye (2023), which randomly selects one channel in a visible image and expands it into a three-channel gray image. The augmented images build an intermediate modality between visible and infrared to enhance training. The OCLR loss for visible modality in V-based training mode can be formulated as follows:

$$\begin{aligned}\mathcal{L}_{OCLR}^v = &\frac{1}{B}\sum_{i=1}^{B}\mathcal{L}_{ce}(\mathbf{P}(\mathbf{f}_i^v|\mathbf{M}^c,\tau),\mathbf{P}(\mathbf{f}_i^a|\tilde{\mathbf{M}}^v,\frac{1}{5}\tau))\\ &+\frac{1}{B}\sum_{i=1}^{B}\mathcal{L}_{ce}(\mathbf{P}(\mathbf{f}_i^a|\mathbf{M}^c,\tau),\mathbf{P}(\mathbf{f}_i^v|\hat{\mathbf{M}}^r,\frac{1}{5}\tau)).\end{aligned} \tag{27}$$

As in Caron et al. (2021); Nassar et al. (2023), we sharpen the target prediction more than the source's to encourage entropy minimization. The setting of the sharpen ratio follows ProtoCon Nassar et al. (2023). The OCLR loss for infrared modality $\mathcal{L}_{OCLR}^r$ in V-based mode can be formulated similarly:

$$\begin{aligned}\mathcal{L}_{OCLR}^r = &\frac{1}{B}\sum_{i=1}^{B}\mathcal{L}_{ce}(\mathbf{P}(\mathbf{f}_i^r|\mathbf{M}^c,\tau),\mathbf{P}(\mathbf{f}_i^r|\tilde{\mathbf{M}}^v,\frac{1}{5}\tau))\\ &+\frac{1}{B}\sum_{i=1}^{B}\mathcal{L}_{ce}(\mathbf{P}(\mathbf{f}_i^r|\mathbf{M}^c,\tau),\mathbf{P}(\mathbf{f}_i^r|\hat{\mathbf{M}}^r,\frac{1}{5}\tau)).\end{aligned} \tag{28}$$

3.4 The Overall Objective Function

The overall training loss can be formulated as follows:

$$\mathcal{L} = \mathcal{L}_{IM} + \mathcal{L}_{CM} + \mathcal{L}_{OCLR}^v + \mathcal{L}_{OCLR}^r, \tag{29}$$

where $\mathcal{L}_{IM}, \mathcal{L}_{CM}, \mathcal{L}_{OCLR}^v$ and $\mathcal{L}_{OCLR}^r$ are described in detail in Sec.3.2 and Sec.3.3.

4 Experiments

4.1 Dataset and Evaluation Protocol

Datasets. Our proposed method is evaluated on two public visible-infrared datasets, namely SYSU-MM01 Wu et al. (2017) and RegDB Nguyen et al. (2017). SYSU-MM01 comprises 395 identities, with 22258 visible and 11909 infrared images captured by indoor and outdoor cameras. The dataset is divided into the training set (296 IDs), the validation set (99 IDs), and the test set (96 IDs). RegDB is a smaller dataset obtained from a pair of aligned visible and infrared cameras Nguyen et al. (2017). It contains 412 identities, where each identity has 10 visible images and 10 infrared images.

Evaluation Metrics. All experiments follow the common evaluation protocols used for VI-ReID Ye et al. (2021b, 2020, 2021a). The evaluation metrics include Cumulative Matching Characteristic (CMC), Mean Average Precision (mAP), and Mean Inverse Negative Penalty (mINP Ye et al. (2021b)). For the SYSU-MM01 dataset, following Wu et al. (2021);

Table 1 Comparison with the state-of-the-art methods on SYSU-MM01 and RegDB. "GUR*" denotes GUR without camera labels. Since our method does not require any camera label information, for fair comparison we do not report the results of GUR with camera labels.

Method	Venue	SYSU-MM01 (Single-shot)						RegDB					
		All-search			Indoor-search			Visible-to-Infrared			Visible-to-Infrared		
		R1	mAP	mINP	R1	mAP	mINP	R1	mAP	mINP	R1	mAP	mINP
Supervised VI-ReID methods													
Zero-PadWu et al. (2017)	ICCV-17	14.80	15.95	–	20.58	26.92	–	17.75	18.90	–	16.63	17.82	–
AlignGANMao et al. (2017)	ICCV-19	42.4	40.7	–	45.9	54.3	–	57.9	53.6	–	56.3	53.4	–
cm-SSFT Lu et al. (2020)	CVPR-20	47.7	54.1	–	–	–	–	72.3	72.9	–	71.0	71.7	–
DDAG Ye et al. (2020)	ECCV-21	54.75	53.02	39.62	61.02	67.98	62.61	69.34	63.46	49.24	68.06	61.80	48.62
AGW Ye et al. (2021b)	TPAMI-21	47.50	47.65	35.30	54.17	62.97	59.23	70.05	66.37	50.19	70.49	65.90	51.24
VCD+VML Tian et al. (2021)	CVPR-21	60.02	58.80	–	66.05	72.98	–	73.2	71.6	–	71.8	70.1	–
CA Ye et al. (2021a)	ICCV-21	69.88	66.89	53.61	76.26	80.37	76.79	85.03	79.14	65.33	84.75	77.82	61.56
MPANet Wu et al. (2021)	CVPR-21	70.58	68.24	–	76.74	80.95	–	82.8	80.7	–	83.7	80.9	–
MAUM Liu et al. (2022)	CVPR-22	71.68	68.79	–	76.97	81.94	–	87.87	85.09	–	86.95	84.34	–
CIFT Li et al. (2022b)	ECCV-22	74.08	74.79	–	81.82	85.61	–	91.96	92.00	–	90.30	90.78	–
DEENZhang and Wang (2023)	CVPR-23	74.7	71.8	–	80.3	83.3	–	91.1	85.1	–	89.5	83.4	–
SEFEL Feng et al. (2023)	CVPR-23	77.12	72.33	–	82.07	82.95	–	91.07	85.23	–	92.18	86.59	–
PartMix Kim et al. (2023)	CVPR-23	77.78	74.62	–	81.52	84.38	–	84.93	82.52	–	85.66	82.27	–
Unsupervised single-modality ReID methods													
SPCL Ge et al. (2020)	NIPS-20	18.37	19.39	10.99	26.83	36.42	33.05	13.59	14.86	10.36	11.70	13.56	10.09
MMT Ge et al. (2020a)	ICLR-20	21.47	21.53	11.50	22.79	31.50	27.66	25.68	26.51	19.56	24.42	25.59	18.66
Cluster-Contrast Dai et al. (2023)	arXiv-21	20.16	22.00	12.97	23.33	34.01	30.88	11.76	13.88	9.94	11.14	12.99	8.99
ICE Chen et al. (2021)	ICCV-21	20.54	20.39	10.24	29.81	38.35	34.32	12.98	15.64	11.91	12.18	14.82	10.6
PPLR Cho et al. (2022)	CVPR-22	11.98	12.25	4.97	12.71	20.81	17.61	10.30	11.94	8.10	10.39	11.23	7.04
ISE Zhang et al. (2022b)	CVPR-22	20.21	18.93	8.54	14.22	24.62	21.74	16.12	16.99	13.24	10.83	13.66	10.71

Table 1 continued

Method	Venue	SYSU-MM01 (Single-shot)						RegDB					
		All-search			Indoor-search			Visible-to-Infrared			Visible-to-Infrared		
		R1	mAP	mINP	R1	mAP	mINP	R1	mAP	mINP	R1	mAP	mINP
Unsupervised VI-ReID methods													
H2H Liang et al. (2021)	TIP-21	25.49	25.16	–	–	–	–	13.91	12.72	–	14.11	12.29	–
H2H(AGW) Liang et al. (2021)	TIP-21	30.15	29.40	–	–	–	–	23.81	18.87	–	–	–	–
H2H(AGW) w/ CMRR Liang et al. (2021)	TIP-21	45.47	47.99	–	–	–	–	35.18	36.46	–	–	–	–
OTLA Wang et al. (2022)	ECCV-22	29.9	27.1	–	29.8	38.8	–	32.9	29.7	–	32.1	28.6	–
ADCA Yang et al. (2022)	MM-22	45.51	42.73	28.29	50.60	59.11	55.17	67.20	64.05	52.67	68.48	63.81	49.62
ADCA(AGW) Yang et al. (2022)	MM-22	50.90	45.70	29.12	51.39	59.82	56.08	66.62	63.47	–	67.29	62.98	–
CHCR (AGW) Pang et al. (2023)	T-CSVT-23	47.72	45.34	–	–	–	–	68.18	63.75	–	69.96	65.87	–
DOTLA (AGW) Cheng et al. (2023b)	MM-23	50.36	47.36	32.40	53.47	61.73	57.35	85.63	76.71	61.58	82.91	74.97	58.60
MBCCM Cheng et al. (2023a)	MM-23	53.14	48.16	32.41	55.21	61.98	57.23	83.79	77.87	65.04	82.82	76.74	61.73
MIMR Pang et al. (2024)	KBS-24	46.56	45.88	–	52.26	60.93	–	68.76	64.33	–	68.76	63.83	–
CCLNet (CLIP) Chen et al. (2023)	MM-23	54.03	50.19	–	56.68	65.12	–	69.94	65.53	–	70.17	66.66	–
PGM(AGW) Wu and Ye (2023)	CVPR-23	57.27	51.78	34.96	56.23	62.74	58.13	69.48	65.41	–	69.85	65.17	–
GUR* Yang et al. (2023)	ICCV-23	60.95	56.99	41.85	64.22	69.49	64.81	73.91	70.23	58.88	75.00	69.94	56.21
Ours	–	**64.77**	**59.23**	**43.46**	**65.34**	**71.46**	**67.83**	**89.95**	**82.09**	**67.29**	**90.78**	**82.25**	**65.38**
Ours w/ CMRR Liang et al. (2021)	Ours + TIP-21	**71.23**	**67.54**	**54.48**	**69.06**	**75.06**	**71.36**	**91.50**	**92.51**	**92.03**	**90.53**	**91.49**	**91.27**

The results of our method are in bold, and the second best results are underlined.

Lu et al. (2020); Mao et al. (2017), we evaluate the proposed method under two search modes: the All-search mode and the Indoor-search mode. In All-search mode, the gallery set contains all visible images from both indoor and outdoor cameras, while in indoor-search mode, the gallery set only contains images from indoor cameras. Additionally, we conduct experiments under both single-shot and multi-shot settings. We repeat the evaluation 10 times with the random split of the query set and the gallery set and report the average performance. As for the RegDB dataset, we evaluate our method on two testing modes: Visible-to-Infrared and Infrared-to-Visible. The dataset is randomly split into training (206 identities) and testing (206 identities) sets. We randomly selected 206 identities for training and the remaining 206 for testing.

Implementation Details. Our proposed method is implemented using PyTorch. we adopt two-stream ResNet50 He et al. (2016) pretrained on ImageNet Deng et al. (2009) as our backbone. In a mini-batch, the number of classes P and instances for each class K are both 12. All the input images are resized to 288×144. The augmentations for images are following Yang et al. (2022). Besides, Linear Transform Generator (LTG Tan et al. (2023)) and ColorJitter augmentations are adopted in the DCL training stage to promote the encoder to extract color-independent features. We train for a total of 90 epochs. The DCL framework is trained in the first 40 epochs, while our proposed framework is trained in the other 50 epochs. We use the Adam optimizer for training the model with weight decay 5e-4. The initial learning rate is set to 3.5e-4 and decays 10 times at the 20th, 60th, and 80th epochs. Following Cheng et al. (2023a); Yang et al. (2022); Wu and Ye (2023), the momentum factor μ is 0.1, and the temperature τ is 0.05, while the maximum distance for DBSCAN is set to 0.6 on SYSU-MM01 and 0.3 on RegDB. The parameter κ for κ-reciprocal nearest neighbors in Eq.1 and Eq.2 is set to 30 following Zhong et al. (2017). The hyper-parameter λ in Eq.10 for OT is set to 25. The trade-off parameter α in Eq.14 and Eq.15 is set to 0.2 and β in Eq.21 is set to 0.7.

4.2 Comparison with State-of-the-Art Methods

To demonstrate the efficiency of our proposed methods, we compare our method with state-of-the-art methods under three relevant settings on SYSU-MM01 and RegDB datasets, *i.e.*, supervised VI-ReID methods, unsupervised single-modality ReID methods, and unsupervised VI-ReID methods. The results are shown in Tab.1 and Tab.2.

Comparison with Supervised VI-ReID Methods. Our proposed method achieves competitive performance with supervised method VCD+VML Tian et al. (2021) on SYSU-MM01, and outperform several supervised methods including Zero-Pad Wu et al. (2017), AlignGAN Mao et al. (2017), cm-SSFT Mao et al. (2017), DDAG Ye et al. (2020) and AGW Ye et al. (2021b). Our method surpasses most supervised VI-ReID methods on the RegDB dataset and demonstrates excellent performance close to state-of-the-art methods. The results reveal the significant developmental potential of unsupervised VI-ReID, yet there remains a considerable performance gap compared to the state-of-the-art supervised VI-ReID methods.

Table 2 Comparison with the state-of-the-art methods under the Multi-shot setting on SYSU-MM01.

Method	SYSU-MM01 (Multi-shot)			
	All-search		Indoor-search	
	R1	mAP	R1	mAP
Supervised VI-ReID methods				
Zero-Pad Wu et al. (2017)	19.13	10.89	24.43	18.86
cmGAN Dai et al. (2018)	31.49	22.27	37.00	32.76
AlignGAN Mao et al. (2017)	51.50	33.90	57.10	45.30
cm-SSFT Lu et al. (2020)	63.40	62.00	73.00	72.40
MPANet Wu et al. (2021)	75.58	62.91	84.22	75.11
FMCNet Zhang et al. (2022a)	73.44	56.06	78.86	63.82
CIFT Li et al. (2022b)	79.74	<u>75.56</u>	<u>88.32</u>	<u>86.42</u>
PartMix Kim et al. (2023)	<u>80.54</u>	69.84	87.99	79.95
Unsupervised VI-ReID methods				
H2H Liang et al. (2021)	30.31	19.12	–	–
DFC Si et al. (2023)	44.12	28.36	–	–
MBCCM Cheng et al. (2023a)	<u>57.73</u>	39.78	62.87	52.80
CHCR Pang et al. (2023)	50.12	<u>42.17</u>	–	–
Ours	**71.35**	**52.18**	**76.99**	**64.03**

The results of our method are in bold, and the second best results are underlined.

Comparison with Conventional Unsupervised Single-Modality ReID Methods. The results in Tab.1 indicate that unsupervised single-modality methods cannot effectively address the large modality-gap in cross-modality scenarios Our method outperforms the state-of-the-art unsupervised single-modality ReID method Cluster-Contrast Dai et al. (2023) by a large margin of 37.23% mAP and 44.61% Rank-1 on SYSU-MM01, emphasizing the necessity of proposing cross-modality label association strategies for VI-ReID.

Comparison with Unsupervised VI-ReID Methods. As reported in Tab.1, our method outperforms the state-of-the-art GUR Yang et al. (2023) by 2.24% mAP, 3.82% Rank-1 on SYSU-MM01 (All-search) under the single-shot setting, and 11.86% mAP, 16.04% Rank-1 on RegDB (Visible-to-infrared). As shown in Tab.2, our method showcases remarkable performance, surpassing the state-of-the-art CHCR Pang et al. (2023) by 10.01% mAP and 21.23% Rank-1 on SYSU-MM01 under the multi-shot setting (All-search). (Note that PGM Wu and Ye (2023), GRU Yang et al. (2023), among others, do not provide results for the multi-shot setting on SYSU-MM01.) The proposed method has

Table 3 Ablation study on individual components of our method on SYSU-MM01 and RegDB.

Index	Method	SYSU-MM01						RegDB					
		All-search			Indoor-search			Visible-to-Infrared			Visible-to-Infrared		
		R1	mAP	mINP	R1	mAP	mINP	R1	mAP	mINP	R1	mAP	mINP
1	Baseline	41.76	38.57	24.32	45.83	54.44	50.00	50.58	48.78	35.62	51.17	47.24	33.80
The following methods are all based on the MULT module													
2	Baseline + $\mathcal{L}_{CM}$	60.02	54.92	39.18	60.66	66.92	62.24	83.74	75.55	58.26	85.29	75.08	55.50
3	Baseline + $\mathcal{L}_{IM}$ + $\mathcal{L}_{CM}$	60.18	55.52	39.86	60.48	67.39	62.94	85.87	76.08	58.05	84.13	74.36	55.03
4	Baseline + $\mathcal{L}_{OCLR}$	54.34	51.03	36.21	55.85	62.68	58.12	85.87	78.72	65.41	87.86	79.48	61.70
5	Baseline + $\mathcal{L}_{IM}$ + $\mathcal{L}_{OCLR}$	62.35	58.22	43.10	64.77	70.89	66.38	88.40	81.20	67.08	88.93	80.27	62.73
6	Baseline + $\mathcal{L}_{CM}$ + $\mathcal{L}_{OCLR}$	62.70	58.25	43.11	64.06	70.50	66.11	87.77	80.07	66.40	88.16	79.31	62.68
7	Baseline + $\mathcal{L}_{IM}$ + $\mathcal{L}_{CM}$ + $\mathcal{L}_{OCLR}$	64.77	59.23	43.46	65.34	71.46	67.83	89.95	82.09	67.29	90.78	82.25	65.38

three main advantages: (1) Unlike H2H Liang et al. (2021) and OTLA Wang et al. (2022), our method does not require additional datasets or camera labels for training. (2) Our MULT associates cross-modality labels from a novel structural consistency perspective and can be applied to other unsupervised cross-domain tasks. (3) Our OCLR is a plug-and-play module that can be employed in other unsupervised memory-based methods to handle label noise.

4.3 Ablation Study

To validate the effectiveness of each component of our method, we conduct ablation experiments on SYSU-MM01 and RegDB datasets, as shown in Tab.3. Our baseline consists of a two-stream ResNet trained under the DCL framework Yang et al. (2022). All experiments are conducted with our MULT module. We do not perform an ablation study on $\mathcal{L}_{IM}$ as it does not address the problem of modality discrepancy.

Effectiveness of MULT. The MULT module delivers a +16.35% mAP and +18.26% Rank-1 improvement by directly incorporating $\mathcal{L}_{CM}$ (see 1^{st} row and 2^{nd} row in Tab.3) and achieves +20.66% mAP and +23.01% Rank-1 improvement when trained within our entire AMIRL framework on SYSU-MM01 (see 1^{st} row and 7^{th} row in Tab.3). To further evaluate the effectiveness of MULT, we compare MULT with other cross-modality label association methods in Tab.4, including PGM Wu and Ye (2023), MBCCM Cheng et al. (2023a), CLU Yang et al. (2023) and DOTLA Cheng et al. (2023b). Specifically, we integrate the above methods into our AMIRL framework. The results demonstrate our MULT provides higher-quality associations than other methods, achieving performance exceeding DOTLA Cheng et al. (2023b) by 4.19% mAP without OCLR, and 3.88% mAP by integrating our OCLR module. The core effectiveness of our MULT lies in the generated soft pseudo-labels containing richer contextual structural information within the feature space, which promotes the network's learning of shared fine-grained features both within clusters and across clusters.

Effectiveness of OCLR. The OCLR module provides a performance gain of +12.46% mAP and +12.58% Rank-1 when directly adding it to our baseline (see 1^{st} row and 4^{th} row in Tab.3). When used in conjunction with our AMIRL framework, the OCLR further improves the performance of +3.71% mAP and +4.59% Rank-1 (see 3^{rd} row and 7^{th} row in Tab.3). The OCLR module significantly mitigates the impact of pseudo-label noise while narrowing the modality discrepancy. OCLR utilizes ever-evolving prototypes for online refinement, avoiding overfitting static noisy labels. We also integrate our OCLR into other methods (see upper and lower in Tab.4) based on our AMIRL framework. The results indicate that our OCLR improves performance when collaborating with other cross-modality association methods. The proposed OCLR has the following advantages compared to the unsupervised single-modality ReID method MMT Ge et al. (2020a): (1) MMT focuses on consistency in predictions from different augmentations of the same instance, which does not consider the cross-modality interactions. Conversely, our OCLR emphasizes consistency in relationships between instances and multi-modality prototypes, which enhances cross-modality interactions and prompts learning modality-shared information. Therefore, the proposed OCLR is more suitable for the USL-VI-ReID task. (2) MMT refines predictions using the moving average model, while OCLR conducts refinement across multi-memories without storing past models, significantly reducing GPU memory consumption.

Effectiveness of AMIRL. Our AMIRL achieves +0.60% mAP improvement without OCLR (see 2^{nd} row and 3^{rd} row in Tab.3) and achieves +2.41% mAP with OCLR (see 6^{th} row and 7^{th} row in Tab.3) compared to leveraging $\mathcal{L}_{CM}$ alone. The results illustrate the importance of incorporating intra-modality contrastive loss $\mathcal{L}_{IM}$ into our framework, especially when training with our OCLR module. Moreover,

Table 4 Comparison with other cross-modality label association methods for unsupervised VI-ReID on SYSU-MM01. All experiments are based on our AMIRL framework (Sec.3.2).

Method	Venue	All-search			Indoor-search		
		R1	mAP	mINP	R1	mAP	mINP
Baseline	–	41.76	38.57	24.32	45.83	54.43	50.00
AMIRL (Sec. 3.2) w/o OCLR (Sec. 3.3)							
w/ PGM Wu and Ye (2023)	CVPR-23	50.85	45.50	30.64	51.31	59.73	55.34
w/ BCCM Cheng et al. (2023a)	MM-23	51.35	46.61	31.97	53.44	61.28	56.66
w/ CLU Yang et al. (2023)	ICCV-23	57.09	48.96	31.88	57.11	64.51	60.32
w/ DOTLA Cheng et al. (2023b)	MM-23	57.17	51.33	36.07	55.16	62.53	57.87
w/ MULT (Ours)	–	60.18	55.52	39.86	60.48	67.39	62.94
AMIRL (Sec. 3.2) w/ OCLR (Sec. 3.3)							
w/ PGM Wu and Ye (2023)	CVPR-23	56.43	52.31	37.00	59.28	66.04	61.55
w/ BCCM Cheng et al. (2023a)	MM-23	58.46	53.69	37.99	58.84	66.56	62.25
w/ CLU Yang et al. (2023)	ICCV-23	61.40	54.18	38.11	62.91	68.32	64.08
w/ DOTLA Cheng et al. (2023b)	MM-23	61.53	55.35	39.96	60.82	67.60	63.06
w/ MULT (Ours)	–	64.77	59.23	43.46	65.34	71.46	67.83

Table 5 Ablation study for the auxiliary memory

Method	All-search			Indoor-search	
	R1	mAP	mINP	R1	mAP
w/ Intra	63.77	58.61	43.04	64.54	71.38
w/ Aux	63.06	58.20	42.65	65.06	71.07
Intra & Aux	64.77	59.23	43.46	65.34	71.46

Table 6 Ablation study for the alternative scheme

Method	All-search			Indoor-search	
	R1	mAP	mINP	R1	mAP
w/ MIRL	64.05	58.44	43.06	64.86	70.90
w/ AMIRL	64.77	59.23	43.46	65.34	71.46

we conduct the ablation experiment for the auxiliary intra-modality memory and the alternative training scheme, as shown in Tab.5 and Tab.6. In Tab.5, "w/ Intra" denotes only intra-modality memory banks $\tilde{\mathbf{M}}^v$ and $\tilde{\mathbf{M}}^r$ are involved in $\mathcal{L}_{IM}$ during training, and "w/ Aux" denotes only auxiliary memory banks $\hat{\mathbf{M}}^r$ and $\hat{\mathbf{M}}^v$ are utilized. The results indicate that the cross-modality pseudo-labels also contribute to intra-modality learning. The improvement is attributed to these two types of pseudo-labels revealing two distinct intra-modality structures in feature space, thus correcting each other mutually. In Tab.6, "MIRL" denotes the contrastive learning framework without the alternative scheme (AMIRL) for V-based training and R-based training. The results demonstrate that the alternative scheme further enhances the performance. This is achieved by alternating the utilization of labels from a single-directional MULT (V2R or R2V) to supervise model training across different epochs, thereby effectively mitigating the issue of pseudo-label inconsistency between V2R and R2V MULT. Compared to a basic framework, MSMA Cheng et al. (2023a), we construct an auxiliary memory bank to learn the structure of cross-modality pseudo-labels, this strategy fully exploits the pseudo-labels from MULT.

Table 7 Comparison with various designs for modeling $\mathbf{S}^{he}$.

Method	SYSU-MM01 (All-search)			SYSU-MM01 (Indoor-search)		
	R1	mAP	mINP	R1	mAP	mINP
Jaccrad	61.08	55.69	39.20	62.63	68.49	63.78
Cosine (k=20)	62.19	57.97	42.38	64.54	70.88	66.57
Cosine (k=30)	61.73	57.66	41.91	64.65	70.59	66.08
OT problem	64.77	59.23	43.46	66.67	71.87	67.20

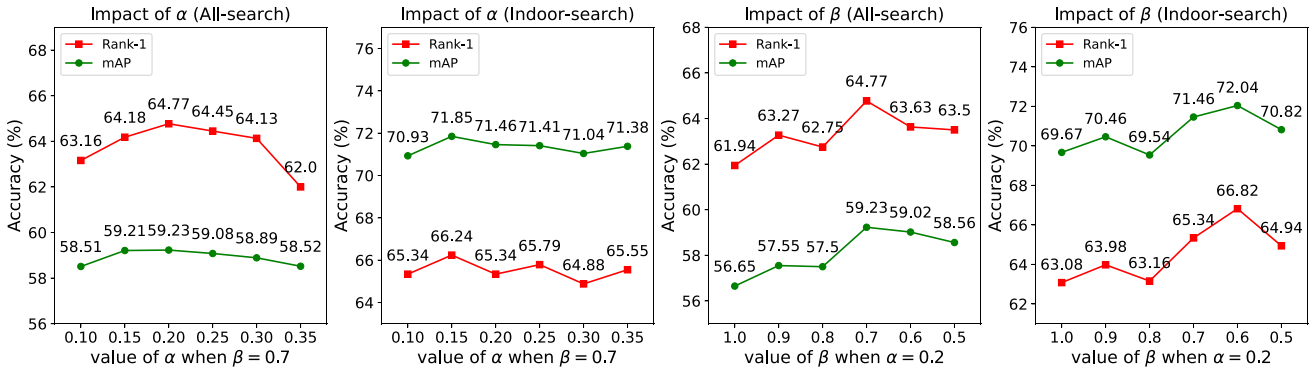

Fig. 4 Parameter analysis of α and β on SYSU-MM01

Furthermore, we propose to alternate V-based training and R-based training. Such an alternative training scheme mitigates the side-effects of noisy labels from the inconsistent pseudo-labels output of the V2R / R2V MULT. our AMIRL makes more effective utilization of the cross-modality pseudo-labels $\tilde{\mathbf{y}}^v$ and $\tilde{\mathbf{y}}^r$, thus is more suitable for our MULT label association algorithm.

4.4 Further Analysis

Analysis for heterogeneous affinities modeling. The Optimal Transport (OT) problem in Eq.3 is formulated to model heterogeneous affinities $\mathbf{S}^{he}$ from the perspective of the transition between instances from two different distributions. The constraint in Eq.3 ensures that N^v visible instances' affinities are uniformly distributed to N^r infrared instances in the transport plan, and vice versa. This design, compared to other alternative designs, prevents degenerated solutions where few infrared instances hold extremely high affinities with the majority of visible instances. We conducted a comparative analysis of various designs for modeling $\mathbf{S}^{he}$, as shown in Tab.7. "Jaccard" denotes the cross-modality Jaccard Similarity-based affinities and "Cosine (k=K)" denotes the cosine-similarity-based knn-affinities, K denotes the number of knn-neighbors for constructing knn-graph. The results confirmed the rationality of our choice.

Parameter Analysis. The proposed method includes two parameters in MULT, *i.e.*, α in Eq.14 and β in Eq.21. α is a trade-off parameter between the self-consistency terms and the heterogeneous consistency terms. β is the parameter to control the smoothness of the soft pseudo-labels from MULT. To investigate the impact of these two parameters, we varied their values, as depicted in Figure 4. We find when $0.15 \leq \alpha \leq 0.3$ and $0.6 \leq \beta \leq 0.7$, the model achieves relatively outstanding performance. We set $\alpha = 0.2$ and $\beta = 0.7$ based on the experiments.

Pseudo-label Accuracy Analysis. To quantify the quality of the pseudo-labels, we calculate the accuracy/recall of positive instance pairs found by the pseudo-labels during training, as shown in Fig. 5 and Fig. 6. We denote the ground-truth hard labels as $y^e(gt)$, where $e = \{v, r\}$ indicates the visible and infrared modality, respectively. We utilize the hard form $y^e = \{\tilde{y}^v, \tilde{y}^r, \hat{y}^v, \hat{y}^r\}$ of the soft pseudo-labels from MULT when quantifying the quality. In Fig. 5, 'Intra-modality visible label accuracy' ($IntraAcc^v$) indicates the accuracy of the visible intra-modality positive pairs found by the cross-modality visible labels $\hat{y}^v$. This can be achieved by computing the proportion of the ground-truth positive pairs within all instance pairs that share the same pseudo-label:

$$IntraAcc^v = \frac{\sum_{i=1}^{N^v}\sum_{j=1}^{N^v}\mathbf{1}(\hat{y}_i^v = \hat{y}_j^v)\mathbf{1}(y_i^v(gt) = y_j^v(gt))}{\sum_{i=1}^{N^v}\sum_{j=1}^{N^v}\mathbf{1}(\hat{y}_i^v = \hat{y}_j^v)}. \tag{30}$$

Similarly, 'Cross-modality visible label accuracy' ($CrossAcc^v$) indicates the accuracy of the cross-modality positive pairs found by $\tilde{\mathbf{y}}^v$ and $\hat{\mathbf{y}}^r$ from the V2R MULT. It can be similarly formulated as follows:

$$CrossAcc^v = \frac{\sum_{i=1}^{N^v}\sum_{j=1}^{N^r}\mathbf{1}(\tilde{y}_i^v = \hat{y}_j^r)\mathbf{1}(y_i^v(gt) = y_j^r(gt))}{\sum_{i=1}^{N^v}\sum_{j=1}^{N^r}\mathbf{1}(\tilde{y}_i^v = \hat{y}_j^r)}. \tag{31}$$

The recall in Fig. 6 can be obtained similarly by computing the proportion of pairs that share the same pseudo-label within all ground-truth positive pairs. We compare our MULT with DOTLA Cheng et al. (2023b), and the results demonstrate that our MULT significantly facilitates the quality of pseudo labels. Simultaneously, the results indicate that as MULT and network training alternate, the network can continuously correct erroneous pseudo-labels.

To further measure the reliability of pseudo-labels derived from our MULT, we employ Fowlkes-Mallows Index (FMI) and Adjusted Rand Index (ARI) metrics to evaluate the pseudo-labels across various association methods, as shown in Tab.8. The FMI score indicates the geometric mean of

Table 8 Comparison with other methods on FMI/ARI

Method	FMI metric (%)				ARI metric (%)			
	RGB ($\tilde{y}^v$)	IR ($\tilde{y}^r$)	$\tilde{y}^v$&$\hat{y}^r$	$\hat{y}^v$&$\tilde{y}^r$	RGB ($\tilde{y}^v$)	IR ($\tilde{y}^r$)	$\tilde{y}^v$&$\hat{y}^r$	$\hat{y}^v$&$\tilde{y}^r$
PGM (CVPR23)	53.57	82.03	56.51	52.05	51.20	81.84	56.30	50.34
BCCM (MM23)	51.99	83.33	60.20	52.61	50.03	83.20	52.29	44.92
OTLA (ECCV22)	58.39	84.40	55.57	51.77	56.40	84.08	55.21	49.29
MULT (Ours)	65.86	86.61	69.32	62.87	63.74	86.48	69.18	61.07

Table 9 Ablation study on CA Ye et al. (2021a) in the proposed OCLR on SYSU-MM01

	SYSU-MM01 (All-search)			SYSU-MM01 (Indoor-search)		
	R1	mAP	mINP	R1	mAP	mINP
Ours w/o OCLR	60.18	55.52	39.18	60.66	66.92	62.24
Ours w/ OCLR (w/o CA)	61.98	57.44	42.05	63.01	69.62	65.42
Ours w/ OCLR (w/ CA)	64.77	59.23	43.46	65.34	71.46	67.83

Table 10 Ablation study on LTG Tan et al. (2023) on SYSU-MM01

Method	RegDB (Visible-to-Infrared)			SYSU-MM01 (All-search)		
	R1	mAP	mINP	R1	mAP	mINP
GUR*(ICCV23,SOTA)	73.91	70.23	58.88	60.95	56.99	41.85
Ours w/o LTG	88.20	81.21	66.24	62.57	57.90	42.25
Ours w/ LTG	89.95	82.09	67.29	64.77	59.23	43.46

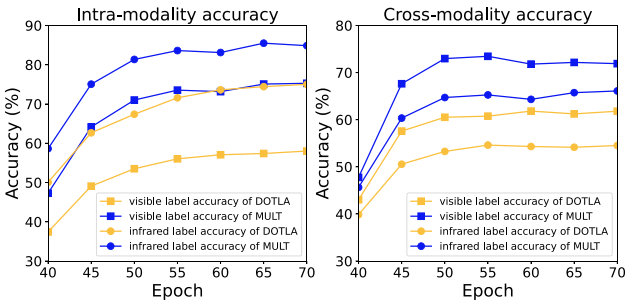

Fig. 5 Accuracy of intra-modality and cross-modality positive pairs found by pseudo-labels on SYSU-MM01

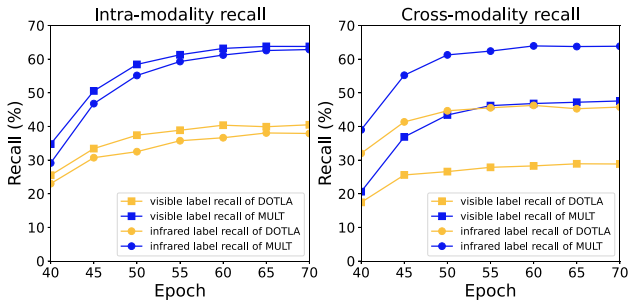

Fig. 6 Recall of intra-modality and cross-modality positive pairs found by pseudo-labels on SYSU-MM01

the accuracy and recall, while the ARI score reflects the degree of overlap between the ground-truth label space and the pseudo-label space. The results illustrate the superiority of our MULT.

Distribution Visualization. We visualize the Euclidean distance distribution of randomly selected 50000 positive and negative visible-infrared pairs, as shown in Fig. 7. By sequentially integrating the modules (*i.e.*, MULT, AMIRL, and OCLR) into the training framework, we observe a convergence of cross-modality positive pairs and a divergence of negative pairs, demonstrating the effectiveness of each component in our framework to address the modality discrepancy. We further visualize the distribution by integrating DOTLA into our framework in Fig. 7d, the results illustrate the superiority of MULT in comparison with DOTLA Cheng et al. (2023b).

T-SNE Visualization. We visualize the t-SNE Van der Maaten and Hinton (2008) map of 7 randomly selected identities, as shown in Fig. 9. In Fig. 9a and b, it is evident that, owing to the precise label associations from MULT, cross-modality positive pairs exhibit increased proximity within the embedding space. Figure 9c and d provide a comparative analysis between our MULT and DOTLA Cheng et al. (2023b). The results reveal the issue of homogeneous inconsistency in DOTLA, where images of the same identities are distributed into distinct clusters within the embedding space (black circles in Fig. 9d). Our MULT effectively addresses this problem (black circles in Fig. 9c) by generating pseudo-labels with low homogeneous consistency. The effectiveness

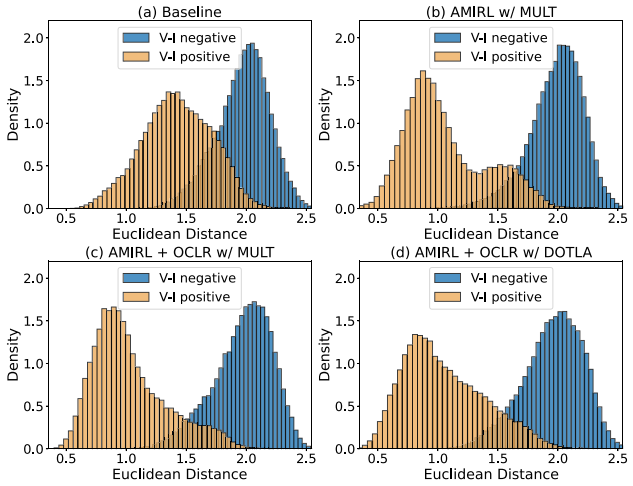

Fig. 7 The visualization of Euclidean distance distribution of randomly selected cross-modality positive and negative pairs

of our OCLR is demonstrated when comparing Fig. 9b and c. Without OCLR, the network struggles to learn compact and discriminative representations due to the inevitable noisy labels (red dotted circles in Fig. 9b). With OCLR, it shows more discriminative features that are well-distributed around the edge of the clusters (green dotted circles in Fig. 9c). The results indicate that our OCLR module is robust to noise while facilitating modality alignment.

Analysis of the data augmentation strategies in the proposed method. Data augmentations are widely-used in VI-ReID, we conduct ablation studies for the data augmentations included in our framework, *i.e.*, CA Ye et al. (2021a) and LTG Tan et al. (2023). CA Ye et al. (2021a) is a commonly-used technology in existing USL-VI-ReID baselines Cheng et al. (2023a); Wu and Ye (2023); Yang et al. (2022); Cheng et al. (2023b); Yang et al. (2023). We conduct ablation studies of the CA augmentation utilized in OCLR, as shown in Tab.9. As shown in Tab.9, the OCLR module is still effective when removing CA, which outperforms the framework without OCLR by 1.92% mAP and 1.8% Rank1. When incorporating the CA-augmented images into our OCLR, the performance further improves by 1.79% mAP and 2.79% Rank1. It can be attributed to the integration of CA bridges different modalities and promotes the model to better learn modality-shared features. We further study the improvement brought by the LTG augmentation, as shown in Tab.10. The results show a slight drop in performance after removing LTG, but it still outperforms the state-of-the-art GUR*. The results also show that LTG further improves the model performance.

Visualization of Ranking List. We visualize some of the ranking lists on the SYSU-MM01 dataset, as shown in Fig. 8. We compare our method with the baseline pretrained under the DCL framework Yang et al. (2022). The results demonstrate the effectiveness of our method in generating high-quality cross-modality supervision signals for training.

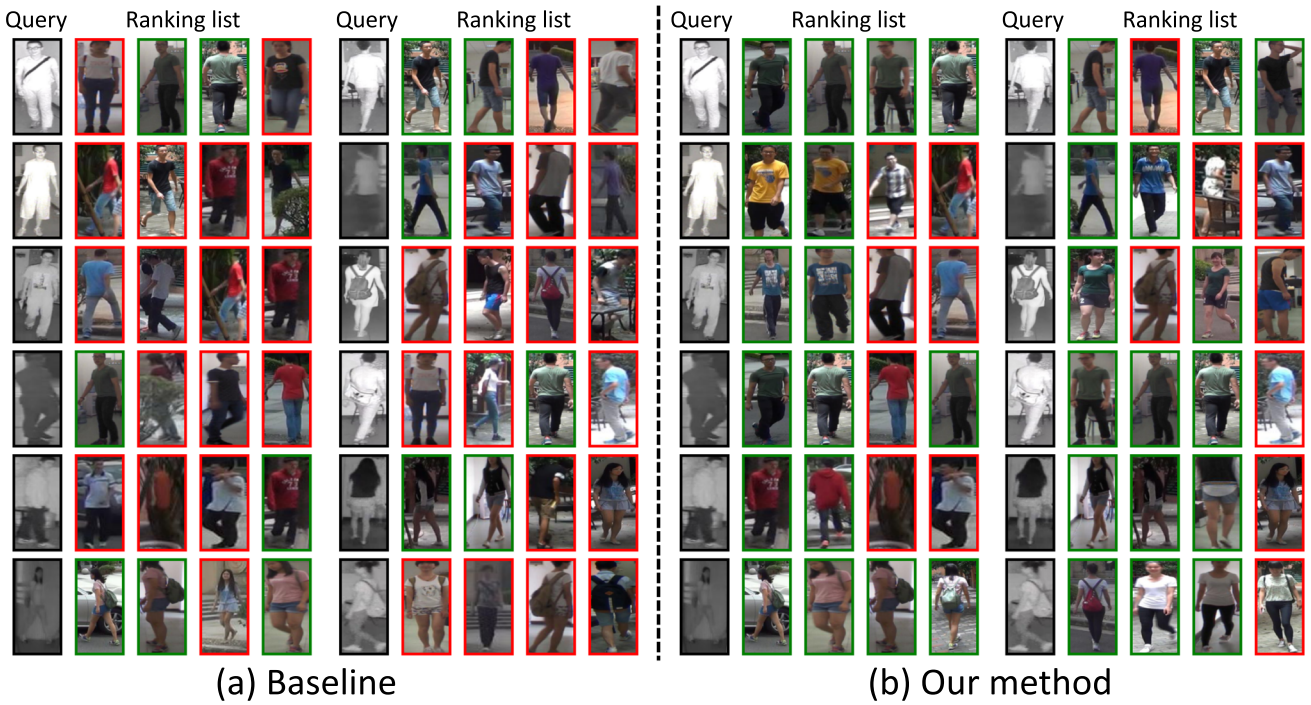

Fig. 8 Visualization of the ranking lists on the SYSU-MM01 dataset. The persons who are different from the query persons are marked with red boxes, while those who are the same as the query are marked with green boxes

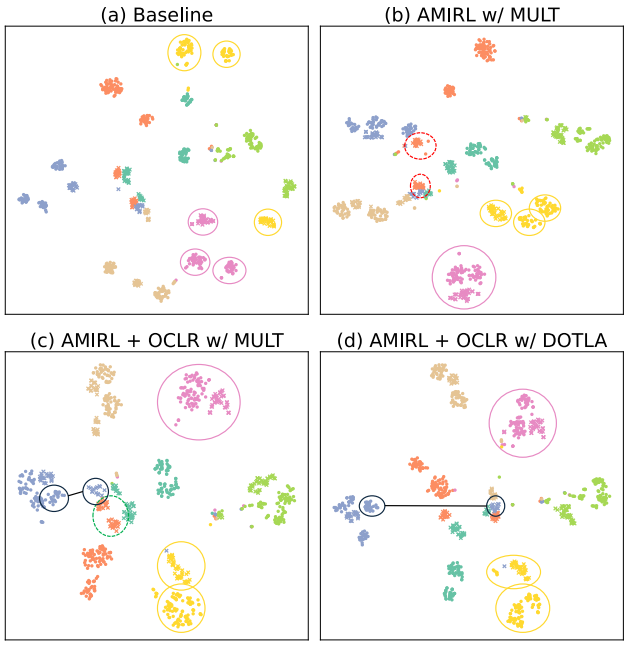

Fig. 9 The t-SNE Van der Maaten and Hinton (2008) visualization of the feature of 7 randomly selected identities. Different colors mean different identities. Circle means the visible modality and cross means the infrared modality

5 Conclusion

In this paper, we propose the Modality-Unified Label Transfer module to establish high-quality cross-modality pseudo-label associations for training from a novel structural consistency perspective. Our MULT lays emphasis on preserving both homogeneous and heterogeneous structure information in pseudo-label space by instance-wise affinities-guided label transfer. To fully exploit the soft pseudo-labels from MULT, an Alternative Modality-Invariant Representation Learning framework is proposed based on both intra-modality and cross-modality contrastive learning. Furthermore, we introduce a straightforward yet effective plug-and-play Online Cross-memory Label Refinement module, which simultaneously mitigates the negative effects of noisy labels and facilitates modality alignment. Extensive experiments have demonstrated that our framework outperforms prior state-of-the-art methods. In the future, we will investigate more robust cross-modality label association methods based on our proposed MULT, which is the core issue of the unsupervised VI-ReID task.

Acknowledgements This work was supported in part by NSFC under Grant NO.62176198, 62036007 and U22A2096, in part by the Key Laboratory of Big Data Intelligent Computing under Grant BDIC-2023-A-004, in part by the Key R&D Program of Shaanxi Province under Grant 2024GX-YBXM.135, in part by the Shaanxi Province Core Technology Research and Development Project under grant 2024QY2-GJHX-11, in part by the Fundamental Research Funds for the Central Universities under GrantQTZX23042.

Data Availability SYSU-MM01: A signed dataset release https://github.com/wuancong/SYSU-MM01agreement must be sent to wuancong@gmail.com or wuanc@mail.sysu.edu.cn to obtain a download link. RegDB: The dataset can be downloaded by submitting a copyright form to http://dm.dongguk.edu/link.html.

References

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 9650–9660.

Chen, D., Xu, D., Li, H., Sebe, N., Wang, X. (2018). Group consistent similarity learning via deep crf for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8649–8658.

Chen, H., Lagadec, B., Bremond, F. (2021). Ice: Inter-instance contrastive encoding for unsupervised person re-identification. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 14960–14969.

Chen, Z., Zhang, Z., Tan, X., Qu, Y., Xie, Y. (2023). Unveiling the power of clip in unsupervised visible-infrared person re-identification. In: Proceedings of the 31st ACM international conference on multimedia, association for computing machinery, New York, NY, USA, MM '23, p 3667–3675, https://doi.org/10.1145/3581783.3612050

Cheng, D., He, L., Wang, N., Zhang, S., Wang, Z., Gao, X. (2023a). Efficient bilateral cross-modality cluster matching for unsupervised visible-infrared person reid. In: Proceedings of the 31st ACM international conference on multimedia, pp 1325–1333.

Cheng, D., Huang, X., Wang, N., He, L., Li, Z., Gao, X. (2023b) Unsupervised visible-infrared person reid by collaborative learning with neighbor-guided label refinement. In: Proceedings of the 31st ACM international conference on multimedia, pp 7085–7093.

Cheng, D., Wang, X., Wang, N., Wang, Z., Wang, X., & Gao, X. (2023). Cross-modality person re-identification with memory-based contrastive embedding. *Proceedings of the AAAI Conference on Artificial Intelligence, 37*(1), 425–432. https://doi.org/10.1609/aaai.v37i1.25116

Cho, Y., Kim, W.J., Hong, S., Yoon, S.E. (2022). Part-based pseudo label refinement for unsupervised person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 7308–7318.

Choi, S., Lee, S., Kim, Y., Kim, T., Kim, C. (2020). Hi-cmd: Hierarchical cross-modality disentanglement for visible-infrared person re-identification. In: proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10257–10266.

Courty, N., Flamary, R., Tuia, D., & Rakotomamonjy, A. (2016). Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39*(9), 1853–1865.

Courty, N., Flamary, R., Tuia, D., Rakotomamonjy, A. (2016). Optimal transport for domain adaptation. 1507.00504.

Dai, P., Ji, R., Wang, H., Wu, Q., & Huang, Y. (2018). Cross-modality person re-identification with generative adversarial training. *In: IJCAI, 1*(3), 6.

Dai, Z., Wang, G., Yuan, W., Liu, X., Zhu, S., & Tan, P. (2023). Cluster contrast for unsupervised person re-identification. *In Proceedings of the Asian Conference on Computer Vision*. https://doi.org/10.48550/arXiv.2103.11568

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition, IEEE, pp 248–255.

Ester, M., Kriegel, H. P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *In: kdd, 96*, 226–231.

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

Ge, Y., Chen, D., Li, H. (2020a). Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. In: International conference on learning representations, https://openreview.net/forum?id=rJlnOhVYPS

Ge, Y., Zhu, F., Chen, D., Zhao, R., et al. (2020). Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. *Advances in Neural Information Processing Systems, 33*, 11309–11321.

Hao, Y., Wang, N., Li, J., & Gao, X. (2019). Hsme: Hypersphere manifold embedding for visible thermal person re-identification. *Proceedings of the AAAI Conference on Artificial Intelligence, 33*, 8385–8392.

He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778.

He, K., Fan, H., Wu. Y., Xie, S., Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9729–9738.

Jia, M., Zhai, Y., Lu, S., Ma, S., Zhang, J. (2020). A similarity inference metric for rgb-infrared cross-modality person re-identification. arXiv preprint arXiv:2007.01504

Kim, M., Kim, S., Park, J., Park, S., Sohn, K. (2023). Partmix: Regularization strategy to learn part discovery for visible-infrared person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 18621–18632.

Kipf, T.N., Welling, M. (2016). Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907

Li, M., Li, C. G., & Guo, J. (2022). Cluster-guided asymmetric contrastive learning for unsupervised person re-identification. *IEEE Transactions on Image Processing, 31*, 3606–3617.

Li, X., Lu, Y., Liu, B., Liu, Y., Yin, G., Chu, Q., Huang. J., Zhu. F., Zhao. R., Yu, N. (2022b). Counterfactual intervention feature transfer for visible-infrared person re-identification. In: European conference on computer vision, Springer, pp 381–398

Liang, W., Wang, G., Lai, J., & Xie, X. (2021). Homogeneous-to-heterogeneous: Unsupervised learning for rgb-infrared person re-identification. *IEEE Transactions on Image Processing, 30*, 6392–6407. https://doi.org/10.1109/TIP.2021.3092578

Lin, X., Li, J., Ma, Z., Li, H., Li, S., Xu, K., Lu, G., Zhang, D. (2022). 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, pp 20973–20982

Lin, Y., Dong, X., Zheng, L., Yan, Y., & Yang, Y. (2019). A bottom-up clustering approach to unsupervised person re-identification. *Proceedings of the AAAI Conference on Artificial Intelligence, 33*, 8738–8745.

Lin, Y., Xie, L., Wu, Y., Yan, C., Tian, Q. (2020). Unsupervised person re-identification via softened similarity learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3390–3399

Liu, J., Sun, Y., Zhu, F., Pei, H., Yang, Y., Li, W. (2022). Learning memory-augmented unidirectional metrics for cross-modality person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 19366–19375

Lu, Y., Wu, Y., Liu, B., Zhang, T., Li, B., Chu, Q., Yu, N. (2020). Cross-modality person re-identification with shared-specific feature transfer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)

Luo, C., Chen, Y., Wang, N., Zhang, Z. (2019). Spectral feature transformation for person re-identification. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 4976–4985

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research, 9*(86), 2579–2605.

Mao, X., Li, Q., Xie, H. (2017). Aligngan: Learning to align cross-domain images with conditional generative adversarial networks. 1707.01400

Nassar, I., Hayat, M., Abbasnejad, E., Rezatofighi, H., Haffari, G. (2023). Protocon: Pseudo-label refinement via online clustering and prototypical consistency for efficient semi-supervised learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 11641–11650

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

Pang, Z., Wang, C., Zhao, L., Liu, Y., Sharma, G. (2023). 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, https://doi.org/10.1109/TCSVT.2023.3310015

Pang, Z., Wang, C., Pan, H., Zhao, L., Wang, J., & Guo, M. (2024). Mimr: Modality-invariance modeling and refinement for unsupervised visible-infrared person re-identification. *Knowledge-Based Systems, 285*, 111350.

Pu, N., Chen, W., Liu, Y., Bakker, E.M., Lew, M.S. (2020). Dual gaussian-based variational subspace disentanglement for visible-infrared person re-identification. In: Proceedings of the 28th ACM international conference on multimedia, pp 2149–2158

Shen, Y., Li, H., Xiao, T., Yi, S., Chen, D., Wang, X. (2018a). Deep group-shuffling random walk for person re-identification. 1807.11178

Shen, Y., Li, H., Yi, S., Chen, D., Wang, X. (2018b). Person re-identification with deep similarity-guided graph neural network. In: Proceedings of the European conference on computer vision (ECCV), pp 486–504

Si, T., He, F., Li, P., Song, Y., & Fan, L. (2023). Diversity feature constraint based on heterogeneous data for unsupervised person re-identification. *Information Processing and Management, 60*(3), 103304.

Tan, L., Zhang, Y., Shen, S., Wang, Y., Dai, P., Lin, X., Wu, Y., Ji, R. (2023). Exploring invariant representation for visible-infrared person re-identification. 2302.00884

Tarvainen, A., Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems 30

Tian, X., Zhang, Z., Lin, S., Qu, Y., Xie, Y., Ma, L. (2021). Farewell to mutual information: Variational distillation for cross-modal person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 1522–1531

Wang, D., Zhang, S. (2020). Unsupervised person re-identification via multi-label classification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10981–10990

Wang, J., Zhang, Z., Chen, M., Zhang, Y., Wang, C., Sheng, B., Qu, Y., Xie, Y. (2022). Optimal transport for label-efficient visible-infrared person re-identification. In: European conference on computer vision, Springer, pp 93–109

Wang, M., Lai, B., Huang, J., Gong, X., & Hua, X. S. (2021). Camera-aware proxies for unsupervised person re-identification. *Proceedings of the AAAI Conference on Artificial Intelligence, 35*, 2764–2772.

Wu, A., Zheng, W.S., Yu, H.X., Gong, S., Lai, J. (2017) Rgb-infrared cross-modality person re-identification. In: Proceedings of the IEEE international conference on computer vision, pp 5380–5389

Wu, J., Liu, H., Su, Y., Shi, W., Tang, H. (2023). Learning concordant attention via target-aware alignment for visible-infrared person re-identification. In: Proceedings of the IEEE/CVF international conference on computer vision (ICCV), pp 11122–11131

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

Wu, Z., Ye, M. (2023). Unsupervised visible-infrared person re-identification via progressive graph matching and alternate learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9548–9558

Yang, B., Ye, M., Chen, J., Wu, Z. (2022). Augmented dual-contrastive aggregation learning for unsupervised visible-infrared person re-identification. In: Proceedings of the 30th ACM international conference on multimedia, association for computing machinery, New York, NY, USA, MM '22, p 2843–2851, https://doi.org/10.1145/3503161.3548198,

' Yang, B., Chen, J., Ye, M. (2023). Towards grand unified representation learning for unsupervised visible-infrared person re-identification. In: Proceedings of the IEEE/CVF international conference on computer vision (ICCV), pp 11069–11079

Ye, M., Shen, J., J Crandall, D., Shao, L., Luo, J. (2020). Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In: Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16, Springer, pp 229–247

Ye, M., Ruan, W., Du, B., Shou, M.Z. (2021a). Channel augmented joint learning for visible-infrared recognition. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 13567–13576

Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., & Hoi, S. C. H. (2001). *2021b* (p. 04193). Deep learning for person re-identification: A survey and outlook.

Yu, H.X., Zheng, W.S., Wu, A., Guo, X., Gong, S., Lai, J.H. (2019). Unsupervised person re-identification by soft multilabel learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2148–2157

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

Zhang, Q., Lai, C., Liu, J., Huang, N., Han, J. (2022a). Fmcnet: Feature-level modality compensation for visible-infrared person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 7349–7358

Zhang, X., Ge, Y., Qiao, Y., Li, H. (2021a). Refining pseudo labels with clustering consensus over generations for unsupervised object re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3436–3445

Zhang, X., Li, D., Wang, Z., Wang, J., Ding, E., Shi, J.Q., Zhang, Z., Wang, J. (2022b) Implicit sample extension for unsupervised person re-identification. In: proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 7369–7378

Zhang, Y., Wang, H. (2023). 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 (CVPR), pp 2153–2162

Zhang, Y., Yan, Y., Lu, Y., Wang, H. (2021b). Towards a unified middle modality learning for visible-infrared person re-identification. In: Proceedings of the 29th ACM international conference on multimedia, pp 788–796

Zhong, Z., Zheng, L., Cao, D., Li, S. (2017). Re-ranking person re-identification with k-reciprocal encoding. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1318–1327

Zhou, D., Bousquet, O., Lal, T., Weston, J., Schölkopf, B. (2003). Learning with local and global consistency. Advances in neural information processing systems 16

Zou, C., Chen, Z., Cui, Z., Liu, Y., Zhang, C. (2023). Discrepant and multi-instance proxies for unsupervised person re-identification. In: Proceedings of the IEEE/CVF international conference on computer vision (ICCV), pp 11058–11068