Unsupervised Visible-Infrared Person Re-identification under Unpaired Settings

Haoyu Yao[1,2,3,4*] Bin Yang[1,2,3*] Wenke Huang[1,2,3] Bo Du[1,2,3] Mang Ye[1,2,3†]
[1] School of Computer Science, Wuhan University, China
[2] National Engineering Research Center for Multimedia Software, Wuhan University, China
[3] Hubei Key Laboratory of Multimedia and Network Communication Engineering, Wuhan University, China
[4] School of Remote Sensing and Information Engineering, Wuhan University, China

https://github.com/USL-VI-ReID/MCL

Abstract

Unsupervised visible-infrared person re-identification (USL-VI-ReID) aims to train a cross-modality retrieval model without labels, reducing the reliance on expensive cross-modality manual annotation. However, existing USL-VI-ReID methods rely on artificially cross-modality paired data as implicit supervision, which is also expensive for human annotation and contrary to the setting of unsupervised tasks. In addition, this full alignment of identity across modalities is inconsistent with real-world scenarios, where unpaired settings are prevalent. To this end, we study the USL-VI-ReID task under unpaired settings, which uses cross-modality unpaired and unlabeled data for training a VI-ReID model. We propose a novel Mapping and Collaborative Learning (MCL) framework. Specifically, we first design a simple yet effective Cross-modality Feature Mapping (CFM) module to map and generate fake cross-modality positive feature pairs, constructing a cross-modal pseudo-identity space for feature alignment. Then, a Static-Dynamic Collaborative (SDC) learning strategy is proposed to align cross-modality correspondences through a collaborative approach, eliminating inter-modality discrepancies across different aspects i.e., *cluster-level and instance-level, in scenarios with cross-modal identity mismatches. Extensive experiments on the conducted SYSU-MM01 and RegDB benchmarks under paired and unpaired settings demonstrate that our proposed MCL significantly outperforms existing unsupervised methods, facilitating USL-VI-ReID to real-world deployment.*

1. Introduction

Person re-identification (ReID) focuses on retrieving specific individuals across non-overlapping cameras [1, 50].

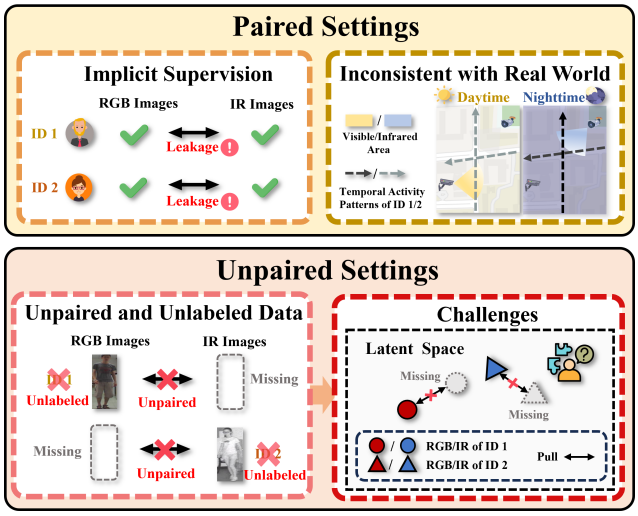


Figure 1. Illustration of the motivation. Previous advanced works for USL-VI-ReID mainly focus on paired settings, which leads to the leakage of supervised information and is inconsistent with the real-world scenarios. Under the underexplored unpaired settings, the cross-modality correspondences are hard to align with unpaired and unlabeled data, which poses a great challenge of how to establish the relation of unpaired samples among different modalities for learning modality-invariant knowledge.

Traditional ReID methods primarily rely on daytime visible-light images, which struggle to extract discriminative features in low-light environments, limiting single-modality ReID deployment in real-world surveillance [13, 26]. To address this problem, the cross-modality visible-infrared person re-identification (VI-ReID) is proposed to identify the same person across a set of visible and infrared images [42]. While supervised VI-ReID methods use plenty of costly cross-modality (visible-infrared) identity labels to learn modality-invariant feature representations [15, 19, 30, 38, 41], the unsupervised VI-ReID (USL-VI-ReID) methods eliminate the need for annotation by generating pseudo-labels [20, 31, 32], which reduce the cost of

*Equal contribution.
†Corresponding author.

expensive cross-modality annotations.

However, existing USL-VI-ReID methods [2–4, 16, 31, 33, 37] operate under the unrealistic assumption of precisely paired cross-modality data, which contradicts real-world scenarios where unpaired settings are more common, as shown in Fig. 1. For instance, a commuter captured by daytime visible cameras may never appear in nighttime infrared surveillance due to social behaviors or hidden activity patterns [14]. More critically, this unrealistic assumption introduces supervised information leakage, violating the unsupervised paradigm and necessitating extensive, costly annotations. To bridge this gap, we introduce a new task: unsupervised visible-infrared person re-identification under unpaired settings, which seeks to train a cross-modality ReID model using unpaired and unlabeled data. When applied to unpaired scenarios, existing USL-VI-ReID methods face two fundamental challenges, as shown in Fig. 1. First, pseudo-label generation depends on cross-modality correspondences, which are absent in unpaired settings, leading to erroneous label assignments. Second, the lack of identity pairs across modalities amplifies the modality gap, hindering effective feature alignment and obstructing the learning of modality-invariant representations.

To overcome these challenges, we construct the first public visible-infrared pedestrian benchmarks under unpaired settings. Based on these benchmarks, we propose a novel Mapping and Collaborative Learning (MCL) framework that establishes cross-modality associations without paired supervision. To handle the absence of paired data, the Cross-modality Feature Mapping (CFM) module is designed to supplement the missing cross-modality paired data while preserving identity consistency. Then, a Static-Dynamic Collaborative (SDC) learning strategy is proposed to achieve unprecedented capability in bridging modality gaps through dual-level alignment. The static learning could establish holistic identity prototypes that capture comprehensive identity characteristics at cluster level, ensuring robust alignment under severe modality mismatches and improving the reliability of cross-modality correspondence. For the dynamic learning strategy, it can reduce the discrepancy between the two modalities by pulling fake samples towards their corresponding actual ones, thereby narrowing the modality gap through instance-level inter-modality alignment. Joint learning of the static and dynamic method forms a collaborative learning strategy, achieving better learning of discriminative and modality-invariant representations for cross-modality retrieval.

The main contributions can be summarized as follows:

- We formally characterize the prevalent unpaired settings encountered in real-world scenarios and introduce the first public visible-infrared pedestrian benchmarks under such conditions. To our knowledge, this work marks the inaugural exploration of unsupervised scenarios featuring cross-modal identity mismatches.
- We propose a novel Mapping and Collaborative Learning (MCL) framework to address the problem of lacking cross-modality paired and labeled data under unpaired settings, establishing a kind of implicit cross-modality associations without paired supervision for learning modality-invariant representations.
- We introduce a straightforward yet effective Cross-modality Feature Mapping (CFM) module that synthesizes positive feature pairs across modalities to achieve robust alignment. Building upon these synthesized pairs, a novel Static-Dynamic Collaborative (SDC) learning strategy is designed to mitigate cross-modality discrepancies at both the cluster and instance levels by leveraging complementary static and dynamic optimization.
- Extensive experiments on two benchmark datasets demonstrate that the proposed framework surpasses existing state-of-the-art USL-VI-ReID methods in unpaired settings, while maintaining competitive performance under paired scenarios.

2. RELATED WORK

2.1. Supervised Visible-Infrared Person ReID

Supervised visible-infrared person re-identification (VI-ReID) has received considerable attention for its applicability in 24-hour surveillance systems. A key challenge is to mitigate the significant intra-class discrepancies between the visible and infrared modalities [29]. Current approaches bridging this cross-modality gap can fall into two categories: image-level and feature-level matching. Image-level methods [25, 36] focus on the generation of cross-modal images to extract modality-invariant features. In contrast, feature-level alignment approaches [9, 17, 18, 28, 35, 40] impose constraints to embed heterogeneous images into a shared feature space. However, these methods rely heavily on extensive cross-modality annotation, which is expensive and time-consuming, making supervised VI-ReID less scalable in real-world deployments.

2.2. Unsupervised Visible-Infrared Person ReID

Unsupervised visible-infrared person re-identification (USL-VI-ReID) faces two key challenges. First, the significant modality gap between visible and infrared data amplifies intra-class variation, making it difficult to consistently generate cross-modality pseudo-labels. Second, the lack of annotated cross-modality identities prevents the learning of modality-invariant representations. Previous work [16, 23, 24, 32, 34, 43, 51] are mainly based on pseudo labels, which establish a bridge with the supervised method. H2H [16] first pioneers a two-stage learning framework, and OTLA [24] introduces an optimal transport strategy for further pseudo-label assignment.

However, both methods rely on external RGB datasets for pre-training, while OTLA also assumes uniform infrared-to-visible label distributions, which is an impractical constraint for real-world scalability. Yang et al. [32] explore cluster-level relationships through cross-modality memory aggregation, but fail to address identity mismatch scenarios. Critically, existing approaches require full cross-modality identity alignment during training, rendering them inapplicable to unpaired data settings.

3. Methodology

3.1. Preliminary

We propose a Mapping and Collaborative Learnin (MCL) framework for USL-VI-ReID under unpaired settings, as shown in Fig. 2. Our MCL has two components including Cross-modality Feature Mapping (CFM) module and Static-Dynamic Collaborative (SDC) learning strategy. In this paper, we follow Augmented Dual-Contrastive Aggregation (ADCA) [32] to establish our baseline, where a dual-path contrastive learning framework with two modality-specific memories to learn intra-modality representations.

To facilitate the description of our method, we first introduce the notations used in this paper. Let $\boldsymbol{X}_m = \{\boldsymbol{x}_m^1, \boldsymbol{x}_m^2, ..., \boldsymbol{x}_m^{N_m}\}$ denote the unlabeled infrared/visible images with N_m instances, where $m \in \{i, v\}$ denotes the infrared and visible modality, respectively. $\boldsymbol{U}_m = \{\boldsymbol{u}_m^1, \boldsymbol{u}_m^2, ..., \boldsymbol{u}_m^{N_m}\}$ represents the corresponding features extracted by the modality-specific feature extractor f_m^θ. $\boldsymbol{q}_m$ is the query instance feature extracted by f_m^θ. $\{\overline{y}_m^k\}_{k=1}^{N_m}$ represents the ground-truth labels of different modalities. In this work, we study the under-explored unpaired settings of USL-VI-ReID, where there are unpaired cross-modality data in the training set, *i.e.*, there exist k, l that satisfy $\overline{y}_i^k \neq \overline{y}_v^l$. In comparison, existing USL-VI-ReID methods are under the traditional label settings where a large number of cross-modality positive pairs are captured in the training data, *i.e.*, $\overline{y}_i^k = \overline{y}_v^l$ for all k, l.

3.2. Cross-modality Feature Mapping

In USL-VI-ReID, the feature extractor f_m^θ is designed to extract discriminative representation $\boldsymbol{u}_m^k = f_m^\theta(\boldsymbol{x}_m^k)$ for image $\boldsymbol{x}_m^k$ to match, as described by:

$$D(\boldsymbol{u}_{m_a}^a, \boldsymbol{u}_{m_p}^p) < D(\boldsymbol{u}_{m_a}^a, \boldsymbol{u}_{m_n}^n), \tag{1}$$

where D denotes a distance function, and $\boldsymbol{u}_{m_a}^a$, $\boldsymbol{u}_{m_p}^p$, and $\boldsymbol{u}_{m_n}^n$ represent the anchor, positive, and negative features extracted from modalities m_a, m_p, and m_n, respectively. The goal is to ensure that the anchor is closer to the positive than to the negative in the feature space. By reducing the distance between cross-modality positive pairs (*i.e.*, $\boldsymbol{u}_{m_a}^a$ and $\boldsymbol{u}_{m_p}^p$ with $m_a \neq m_p$), the discrepancy can be eliminated across modalities, which facilitates the learning of discriminative features.

However, in unpaired settings, the lack of cross-modality positive pairs in the training data poses a significant challenge. Given an anchor feature $\boldsymbol{u}_{m_a}^a$, it is difficult to find a corresponding positive sample $\boldsymbol{u}_{m_p}^p$ from a different modality, which hinders the ability to model intra-class variations across modalities. Therefore, establishing the relation between unpaired cross-modal data is paramount.

To handle this situation, we substitute the unavailable actual cross-modality positive pair $(\boldsymbol{u}_{m_a}^a, \boldsymbol{u}_{m_p}^p)$ with a synthetic or fake positive pair $(\boldsymbol{u}_{m_a}^a, \hat{\boldsymbol{u}}_{m_p}^p)$. The fake feature $\hat{\boldsymbol{u}}_{m_p}^p$ is generated through a modality-aware transformation:

$$\hat{\boldsymbol{u}}_{m_p}^p = \text{Mapping}(\boldsymbol{u}_{m_a}^a, m_p), \tag{2}$$

where Mapping denotes a function that projects the feature $\boldsymbol{u}_{m_a}^a$ from its source modality m_a into the feature space of the target modality m_p.

To achieve the transformation, we introduce a cross-modality feature mapping module comprising two modality-specific mappers, each of which estimates the modality-dependent moment statistics $\{\boldsymbol{\mu}_m, \boldsymbol{\sigma}_m^2\}$ of feature distributions. For a given set of features from modality m, the associated mapper derives its mean and variance by:

$$\boldsymbol{\mu}_m = \frac{1}{N_m}\sum_{k=1}^{N_m}\boldsymbol{u}_m^k, \quad \boldsymbol{\sigma}_m^2 = \frac{1}{N_m}\sum_{k=1}^{N_m}\left(\boldsymbol{u}_m^k - \boldsymbol{\mu}_m\right)^2, \tag{3}$$

where $\boldsymbol{\mu}_m$ and $\boldsymbol{\sigma}_m$ are the modality-specific mean and variance, respectively.

In the cross-modality feature mapping process, a feature $\boldsymbol{u}_m^k$ of modality m is transformed to a fake feature in the modality-specific distribution of a different modality m' by:

$$\hat{\boldsymbol{u}}_{m'}^k = \text{Mapping}(\boldsymbol{u}_m^k, m') = \gamma\frac{\boldsymbol{u}_m^k - \boldsymbol{\mu}_{m'}}{\sqrt{\boldsymbol{\sigma}_{m'}^2 + \epsilon}} - \zeta, \tag{4}$$

where γ and ζ are scaling and shifting parameters of m'-specific mapper, which are learned during training. ϵ is a small constant to ensure stability.

The CFM module applies a cross-modality affine transformation to align source features with the target distribution by matching mean and variance. We precisely estimate the distributions of both modalities and use the transformation to map features. Since the actual feature $\boldsymbol{u}_m^k$ and mapped feature $\hat{\boldsymbol{u}}_{m'}^k$ are encoded from the same image $\boldsymbol{x}_m^k$, we regard they share the same identity $(\hat{y}_{m'}^k = y_m^k)$ and respectively construct an original modality-specific space and a cross-modal pseudo-identity space for further alignment.

3.3. Static-Dynamic Collaborative Learning

The CFM module complements the missing cross-modality unpaired data by estimating modality-specific feature distributions and transforming them into different modalities.

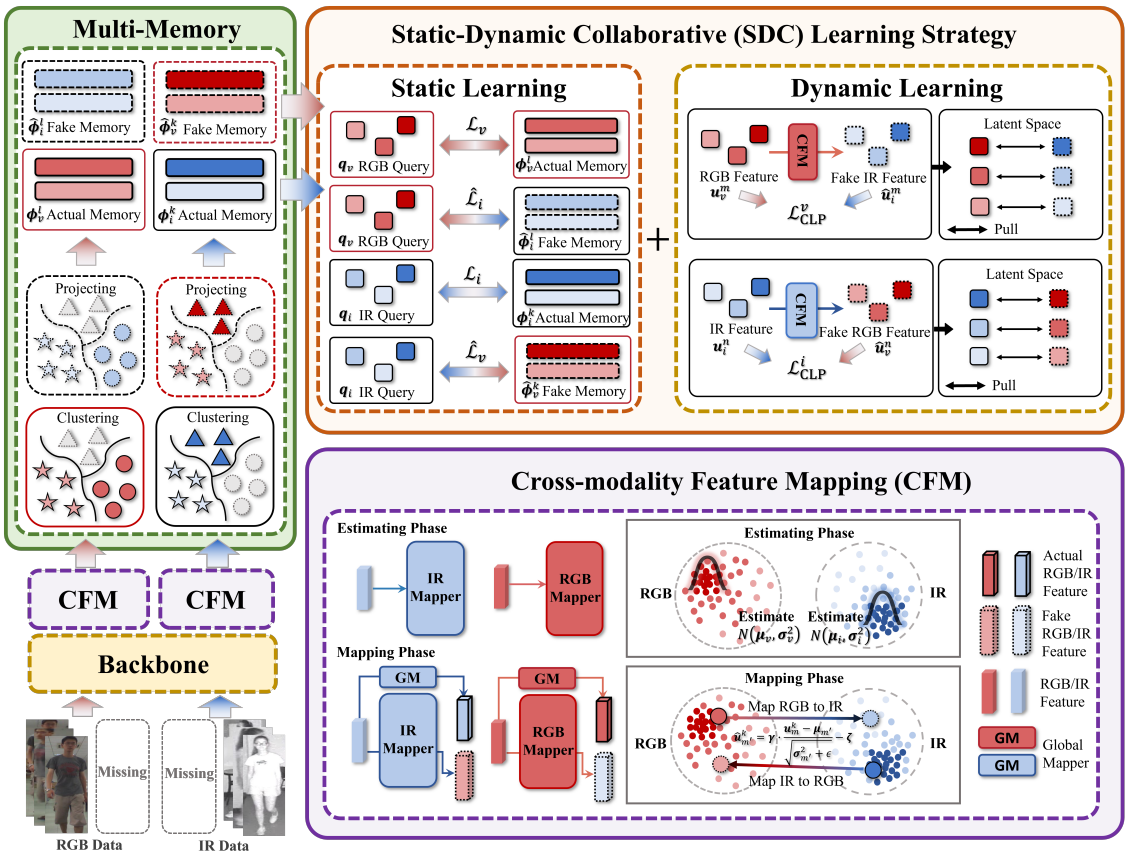


Figure 2. Illustration of mapping and collaborative learning framework. The framework integrates two core components: a Cross-Modality Feature Mapping (CFM) module and a Static-Dynamic Collaborative (SDC) learning strategy. The CFM synthesizes cross-modality positive feature pairs by estimating modality-specific distributions. Concurrently, the SDC strategy addresses inter-modality discrepancies through a twofold alignment: (1) Static Learning employs multi-memory banks to derive reliable cross-modality correspondences by learning holistic representations, establishing robust cluster-level associations. (2) Dynamic Learning utilizes a cross-modality label-preserving loss (CLP) to bridge substantial modality gaps, achieving fine-grained instance-level alignment.

However, it does not use any features prior to and after the mapping process to align cross-modal correspondences. To associate fake cross-modality positive feature pairs and effectively eliminate inter-modality discrepancies, we propose a static-dynamic collaborative learning framework to bridge the gap between the visible and infrared modalities both at cluster-level and instance-level.

Static Learning. We note that the existing methods typically rely on a single memory to represent individual characteristics and establish cross-modality correspondences. However, a single memory may not capture all individual nuances in scenarios with cross-modal identity mismatches, which naturally leads to poor cross-modality correspondences. Therefore, we design a static learning method to obtain reliable cross-modality correspondences by clustering actual samples and fake cross-modality samples as multi-memory banks, jointly conducting the contrastive learning in the original modality-specific space and the cross-modal pseudo-identity space.

At the beginning of each training epoch, all infrared and visible features $\boldsymbol{U}_i$ and $\boldsymbol{U}_v$ are firstly clustered to generate pseudo labels, where each cluster's representations $\{\phi_i^1, ..., \phi_i^K\}$ and $\{\phi_v^1, ..., \phi_v^L\}$ of infrared and visible features are stored in infrared and visible memory dictionaries. This process can be written as:

$$\boldsymbol{\phi}_i^k = \frac{1}{|\mathcal{H}_i^k|} \sum_{\boldsymbol{u}_i^n \in \mathcal{H}_i^k} \boldsymbol{u}_i^n, \boldsymbol{\phi}_v^l = \frac{1}{|\mathcal{H}_v^l|} \sum_{\boldsymbol{u}_v^m \in \mathcal{H}_v^l} \boldsymbol{u}_v^m, \quad (5)$$

where $\mathcal{H}_{i(v)}^k$ is the k-th cluster set in infrared or visible modality, and $|\cdot|$ is the number of instances per cluster.

Then, these actual features $\boldsymbol{U}_i$ and $\boldsymbol{U}_v$ are transformed into corresponding fake cross-modality features $\hat{\boldsymbol{U}}_v$ and $\hat{\boldsymbol{U}}_i$ through CFM. While this mapping process alters the feature distributions, it preserves the original identity information. To maintain consistency in the number of clusters, DBSCAN is applied solely to the actual features for generating pseudo identity labels. Consequently, the cluster representations of the fake cross-modality features are denoted as $\{\hat{\phi}_v^1, \ldots, \hat{\phi}_v^K\}$ and $\{\hat{\phi}_i^1, \ldots, \hat{\phi}_i^L\}$, which constitute two additional memory banks in the cross-modal pseudo-

identity space. This process is formally defined as:

$$\hat{\boldsymbol{\phi}}_v^k = \frac{1}{\left|\hat{\mathcal{H}}_v^k\right|} \sum_{\hat{\boldsymbol{u}}_v^n \in \hat{\mathcal{H}}_v^k} \hat{\boldsymbol{u}}_v^n, \hat{\boldsymbol{\phi}}_i^l = \frac{1}{\left|\hat{\mathcal{H}}_i^l\right|} \sum_{\hat{\boldsymbol{u}}_i^m \in \hat{\mathcal{H}}_i^l} \hat{\boldsymbol{u}}_i^m, \quad (6)$$

where $\hat{\mathcal{H}}_{i(v)}^k$ denotes the k-th cluster set in the mapped infrared or visible modality, and $|\cdot|$ indicates the number of instances per cluster.

During training, we sample P person identities and Z instances for each identity from each modality training set. Then, we use a batch of infrared and visible queries to update the actual memories by a momentum updating strategy:

$$\boldsymbol{\phi}_i^{k(\delta)} \leftarrow \beta \boldsymbol{\phi}_i^{k(\delta-1)} + (1-\beta)\boldsymbol{q}_i, \quad (7)$$

$$\boldsymbol{\phi}_v^{l(\delta)} \leftarrow \beta \boldsymbol{\phi}_v^{l(\delta-1)} + (1-\beta)\boldsymbol{q}_v, \quad (8)$$

The memories of fake cross-modality features are updated by a similar momentum strategy:

$$\hat{\boldsymbol{\phi}}_v^{k(\delta)} \leftarrow \beta \hat{\boldsymbol{\phi}}_v^{k(\delta-1)} + (1-\beta)\boldsymbol{q}_i, \quad (9)$$

$$\hat{\boldsymbol{\phi}}_i^{l(\delta)} \leftarrow \beta \hat{\boldsymbol{\phi}}_i^{l(\delta-1)} + (1-\beta)\boldsymbol{q}_v, \quad (10)$$

where β is the momentum factor and δ is the iteration step.

In each iteration, the feature extractors are jointly updated by a multi ClusterNCE [7] loss, including the actual infrared loss $\mathcal{L}_i$, actual visible loss $\mathcal{L}_v$, fake infrared loss $\hat{\mathcal{L}}_i$, and fake visible loss $\hat{\mathcal{L}}_v$ by the following equations:

$$\mathcal{L}_i = -\log \frac{\exp\left(\boldsymbol{q}_i \cdot \boldsymbol{\phi}_i^+/\tau\right)}{\sum_{k=1}^K \exp\left(\boldsymbol{q}_i \cdot \boldsymbol{\phi}_i^k/\tau\right)}, \quad (11)$$

$$\mathcal{L}_v = -\log \frac{\exp\left(\boldsymbol{q}_v \cdot \boldsymbol{\phi}_v^+/\tau\right)}{\sum_{l=1}^L \exp\left(\boldsymbol{q}_v \cdot \boldsymbol{\phi}_v^l/\tau\right)}, \quad (12)$$

$$\hat{\mathcal{L}}_i = -\log \frac{\exp\left(\boldsymbol{q}_v \cdot \hat{\boldsymbol{\phi}}_i^+/\tau\right)}{\sum_{l=1}^L \exp\left(\boldsymbol{q}_v \cdot \hat{\boldsymbol{\phi}}_i^l/\tau\right)}, \quad (13)$$

$$\hat{\mathcal{L}}_v = -\log \frac{\exp\left(\boldsymbol{q}_i \cdot \hat{\boldsymbol{\phi}}_v^+/\tau\right)}{\sum_{k=1}^K \exp\left(\boldsymbol{q}_i \cdot \hat{\boldsymbol{\phi}}_v^k/\tau\right)}, \quad (14)$$

where $\boldsymbol{\phi}_i^+$ and $\boldsymbol{\phi}_v^+$ are the positive representation vector of the actual infrared and visible cluster, respectively, corresponding to the pseudo label of the query. $\hat{\boldsymbol{\phi}}_i^+$ and $\hat{\boldsymbol{\phi}}_v^+$ are the positive feature vector of the fake infrared and visible cluster. The τ is a temperature hyper-parameter.

Certainly, four types of ClusterNCE [5] loss are designed to learn discriminative representation:

$$\mathcal{L}_{\mathrm{SL}} = \mathcal{L}_i + \mathcal{L}_v + \hat{\mathcal{L}}_i + \hat{\mathcal{L}}_v. \quad (15)$$

Dynamic Learning. The generated cross-modality samples are useful in bridging significant inter-modality discrepancies by including the abundant information of missing identities. However, it is expected that these samples will not change significantly, as they share the same identity with the original samples, which should be aligned during the learning. Inspired by ISE [45], we further propose a Cross-modality Label Preserving (CLP) loss to enforce the fake cross-modality features to be close to their corresponding actual features, which facilitates the learning of cross-modality representations at the instance level. During training, the CLP loss enforces identity consistency between the actual and mapped features, allowing CFM to transfer cross-modal knowledge and capture the complex, non-linear modality discrepancies.

Certainly, two types of CLP loss function are introduced in the dynamic learning method:

$$\mathcal{L}_{\mathrm{CLP}}^i = -\log \frac{\exp(\mathrm{sim}(\boldsymbol{u}_i \cdot \hat{\boldsymbol{u}}_v^+)/\tau')}{\sum_{c=1}^{C_i} \exp(\mathrm{sim}(\boldsymbol{u}_i \cdot \hat{\boldsymbol{u}}_v^{c-})/\tau')}, \quad (16)$$

$$\mathcal{L}_{\mathrm{CLP}}^v = -\log \frac{\exp(\mathrm{sim}(\boldsymbol{u}_v \cdot \hat{\boldsymbol{u}}_i^+)/\tau')}{\sum_{c=1}^{C_v} \exp(\mathrm{sim}(\boldsymbol{u}_v \cdot \hat{\boldsymbol{u}}_i^{c-})/\tau')}, \quad (17)$$

where C_i and C_v are the cluster number of a mini-batch in infrared and visible modality, respectively. $\mathrm{sim}(\cdot)$ denotes the cosine similarity. $\hat{\boldsymbol{u}}_v^+$ is the hardest positive fake visible sample for a given infrared feature $\boldsymbol{u}_i$, and $\hat{\boldsymbol{u}}_v^{c-}$ is the hardest negative one in the c-th fake visible cluster of a mini-batch. $\hat{\boldsymbol{u}}_i^+$ and $\hat{\boldsymbol{u}}_i^{c-}$ are defined similarly to above. τ' is a temperature hyper-parameter.

Consequently, the two cross-modality label preserving loss functions are combined in the dynamic learning method to learn distinctive representations:

$$\mathcal{L}_{\mathrm{DL}} = \mathcal{L}_{\mathrm{CLP}}^i + \mathcal{L}_{\mathrm{CLP}}^v. \quad (18)$$

Collaborative Learning. The cluster-level static learning method and instance-level dynamic learning method jointly forms the static-dynamic collaborative learning strategy, which aims to achieve better learning of discriminative and modality-invariant representations in the absence of cross-modality identity pairs.

The overall loss for training the model is defined by the following equation:

$$\mathcal{L}_{overall} = \mathcal{L}_{\mathrm{SL}} + \eta \mathcal{L}_{\mathrm{DL}}, \quad (19)$$

where η is the loss weight, balancing two loss functions.

4. EXPERIMENTS

4.1. Datasets and Evaluation Protocol

Paired Datasets. To validate the proposed framework under paired settings, we conduct comprehensive evaluations

			SYSU-MM01 (unpaired settings)						RegDB (unpaired settings)					
			All Search			Indoor Search			Visible to Infrared			Infrared to Visible		
α value	Methods	Venue	$r1$	mAP	mINP	$r1$	mAP	mINP	$r1$	mAP	mINP	$r1$	mAP	mINP
0.25	ADCA [32]	MM-22	41.80	37.54	22.04	47.00	56.24	51.79	50.82	43.72	42.63	64.22	55.90	44.29
	OTPA [47]	arXiv-24	49.15	44.75	30.62	51.70	60.82	57.23	51.37	50.30	44.60	66.27	65.82	64.97
	MMM [20]	ECCV-24	54.68	47.35	35.42	55.19	63.76	61.21	62.27	55.09	44.34	76.69	69.93	66.44
	PCAL[37]	TIFS-25	48.84	41.30	26.05	50.48	62.63	59.72	59.33	56.26	45.73	77.18	75.24	63.83
	N-ULC [21]	AAAI-25	53.70	49.69	33.56	57.39	65.00	60.64	60.76	56.82	42.91	79.35	76.06	60.40
	MCL (Ours)	-	**58.57**	**57.74**	**45.45**	**65.72**	**71.84**	**68.30**	**65.97**	**63.53**	**55.84**	**84.31**	**78.67**	**67.28**
0.5	ADCA [32]	MM-22	34.34	32.21	18.76	36.59	47.07	43.23	40.28	32.60	34.93	52.94	47.48	37.05
	OTPA [47]	arXiv-24	48.80	45.64	31.01	49.19	58.74	54.83	40.58	41.52	32.44	59.26	58.68	57.13
	MMM [20]	ECCV-24	44.14	42.94	30.71	54.25	60.17	52.49	52.90	43.91	36.51	63.91	58.17	58.47
	PCAL[37]	TIFS-25	46.24	37.62	22.03	45.39	52.73	46.42	46.48	47.88	36.05	64.21	66.30	50.41
	N-ULC [21]	AAAI-25	49.36	45.12	29.17	51.59	60.24	55.89	50.24	45.96	34.61	70.31	62.84	54.74
	MCL (Ours)	-	**54.29**	**53.98**	**42.64**	**61.71**	**68.45**	**64.40**	**60.78**	**59.69**	**52.91**	**80.36**	**73.02**	**64.25**
0.75	ADCA [32]	MM-22	32.66	30.61	17.98	37.02	47.59	43.79	35.80	28.71	27.96	47.28	43.92	34.70
	OTPA [47]	arXiv-24	33.65	30.50	16.94	45.3	53.32	50.17	31.48	35.60	29.52	51.98	49.25	47.62
	MMM [20]	ECCV-24	36.19	34.60	23.56	43.91	50.60	45.20	41.73	38.64	32.75	54.79	47.90	46.16
	PCAL[37]	TIFS-25	30.74	29.83	20.54	40.60	47.42	39.34	38.10	37.09	33.54	59.80	59.61	47.20
	N-ULC [21]	AAAI-25	42.56	39.44	27.06	55.03	63.11	58.80	43.38	39.27	30.06	66.72	57.51	44.38
	MCL (Ours)	-	**52.16**	**51.57**	**39.57**	**59.63**	**66.46**	**62.24**	**55.30**	**52.79**	**48.26**	**74.19**	**69.37**	**60.58**
1.0	ADCA [32]	MM-22	23.48	23.23	12.53	28.39	37.40	32.80	26.18	15.93	14.72	33.10	28.43	24.95
	OTPA [47]	arXiv-24	20.53	19.46	10.32	36.18	43.54	40.02	19.20	18.03	19.70	39.74	38.42	35.11
	MMM [20]	ECCV-24	30.81	27.53	17.92	36.19	40.60	38.38	30.98	27.10	20.03	40.26	33.54	30.12
	PCAL[37]	TIFS-25	28.95	25.31	18.54	32.10	38.48	36.20	28.74	25.90	18.68	37.56	29.69	27.24
	N-ULC [21]	AAAI-25	34.40	31.83	25.77	40.36	43.70	43.86	34.71	28.06	25.48	42.83	32.40	28.92
	MCL (Ours)	-	**45.98**	**44.95**	**32.85**	**51.53**	**59.06**	**54.54**	**43.18**	**39.29**	**36.95**	**58.37**	**49.40**	**47.21**

Table 1. The comparison with the state-of-the-art methods on SYSU-MM01 and RegDB under unpaired settings. It contains four settings, *i.e.*, the unpaired ratio α ranges from 0.25 to 1.0 in increments of 0.25. Rank-1 accuracy(%), mAP (%) and mINP (%) are reported.

on two benchmark visible-infrared person re-identification datasets: SYSU-MM01 [29] and RegDB [41]. The SYSU-MM01 dataset is a large-scale cross-modality collection of 22,257 visible images and 11,909 near-infrared images captured in indoor and outdoor environments by 4 visible cameras and 2 infrared cameras. In contrast, the RegDB dataset is a smaller and less demanding dataset with 412 different person identities, where each identity contains 10 visible and 10 infrared image pairs. This thermal-infrared dataset is collected using an aligned pair of cameras (one visible and one infrared), which has a comparatively lower environmental complexity than SYSU-MM01.

Unpaired Datasets. To evaluate the effectiveness of our model in unpaired settings, experiments are conducted on the modified SYSU-MM01 and RegDB datasets. It is noted that existing datasets lack cross-modality unpaired training data, thus necessitating adjustments to the existing benchmarks. In order to systematically analyse the impact of unpaired settings, we introduce a hyper-parameter α that controls the proportion of unpaired identities between the visible and infrared modalities. Unlike OTPA [47], which reduces training data by dropping identities, our setup preserves the overall dataset size. Instead of adjusting the overlap ratio by selecting modality-specific subsets, we randomly replace a portion of the visible identities with external images. Specifically, the visible images of selected identities are replaced with images from three visible ReID datasets(*i.e.*, Market-1501 [49], MSMT17 [27] and LLCM [46]), while maintaining a roughly consistent number of images per identity before and after replacement. Details of the replacement process are provided in the **supplementary materials**. This design mitigates the potential effects of data reduction. The test sets and evaluation protocols remain unchanged throughout the experiments.

Evaluation Protocol. On the two benchmark datasets and their original counterpart, we follow the popular protocols [39] for evaluation, where cumulative match characteristic (CMC), mean average precision (mAP), and mean inverse negative penalty (mINP) [41] are adopted. For SYSU-MM01 under paired and unpaired settings, we adopt all-search and indoor-search evaluation modes. For RegDB in paired and unpaired settings, we evaluate our method in the two test modes, including thermal to visible and visible to thermal, and we strictly follow existing methods to perform ten trials of the gallery set selection [40], and calculate the average performance.

4.2. Implementation Details

MCL is implemented in the PyTorch platform. Our work is based on the Augmented Dual-Contrastive Aggregation [32] and incorporates the feature extractor from TransReID [10] as the backbone network. In each mini-batch, 16 identities are selected per modality, with each identity comprising 16 instances. We resize input images to 288 $\times$ 144 pixels for training. Standard data augmentation techniques,

			SYSU-MM01						RegDB					
			All Search			Indoor Search			Visible to Infrared			Infrared to Visible		
	Methods	Venue	$r1$	mAP	mINP	$r1$	mAP	mINP	$r1$	mAP	mINP	$r1$	mAP	mINP
Supervised	AGW [41]	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
	DFLN-ViT [48]	TMM-22	59.84	57.70	-	62.13	69.03	-	92.10	82.11	-	91.21	81.62	-
	PartMix [12]	CVPR-23	77.78	74.62	-	81.52	84.38	-	85.66	82.27	-	84.93	82.52	-
	MUN [44]	ICCV-23	76.24	73.81	-	79.42	82.06	-	95.19	87.15	-	91.86	85.01	-
	SAAI [8]	ICCV-23	75.90	77.03	-	83.20	88.01	-	91.07	91.45	-	92.09	92.01	-
	YYDS [6]	arXiv-24	85.54	81.64	-	89.13	91.00	-	-	-	-	90.20	83.50	-
	TVI-LFM [11]	NIPS-24	84.90	81.47	70.85	89.06	90.78	88.39	-	-	-	91.38	85.92	72.73
Unsupervised	H2H* [16]	TIP-21	30.15	29.40	-	-	-	-	23.81	18.87	-	-	-	-
	OTLA* [24]	ECCV-22	29.90	27.10	-	29.80	38.80	-	32.90	29.70	-	32.10	28.60	-
	ADCA [32]	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
	PGM [31]	CVPR-23	57.27	51.78	34.96	56.23	62.74	58.13	69.48	65.41	-	69.85	65.17	-
	DOTLA* [4]	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 [3]	MM-23	53.14	48.16	32.41	55.21	61.98	57.13	83.79	77.87	65.04	82.82	76.74	61.73
	CCLNet [2]	MM-23	54.03	50.19	-	56.68	65.12	-	69.94	65.53	-	70.17	66.66	-
	GUR† [33]	ICCV-23	60.95	56.99	41.85	64.22	69.49	64.81	73.91	70.23	8.88	75.00	69.94	56.21
	SCA-RCP [15]	TKDE-24	51.41	48.52	33.56	56.77	64.19	59.25	85.59	79.12	-	82.41	75.73	-
	MMM [20]	ECCV-24	61.60	57.90	-	64.40	70.40	-	89.70	80.50	-	85.80	77.00	-
	PCAL [37]	TIFS-25	54.39	51.95	38.09	59.69	66.72	62.44	86.43	82.51	72.33	86.21	81.23	68.71
	N-ULC [21]	AAAI-25	61.81	58.92	45.01	67.04	73.08	69.42	88.75	82.14	68.75	88.17	81.11	66.05
	MCL (Ours)	-	**62.95**	**62.71**	**50.63**	**67.81**	**74.19**	**70.82**	**89.83**	**83.12**	**72.86**	**88.64**	**82.04**	**69.12**

Table 2. The comparison with the state-of-the-art methods on SYSU-MM01 and RegDB under paired settings. It contains two groups, *i.e.*, unsupervised VI-ReID methods and supervised VI-ReID methods. $*$ means the model is pre-trained on an extra labeled visible dataset. GUR† denotes the results without camera information. Rank-1 accuracy(%), mAP (%) and mINP (%) are reported.

	Components				SYSU-MM01* (All Search)			SYSU-MM01* (Indoor Search)			RegDB* (Visible to Infrared)		
Index	Baseline	CFM	SL	DL	$r1$	mAP	mINP	$r1$	mAP	mINP	$r1$	mAP	mINP
1	✔				47.14	47.29	34.63	55.39	63.10	59.43	49.28	48.6	43.93
2	✔	✔	✔		51.65	52.48	40.08	58.38	66.87	63.66	56.39	54.52	49.70
3	✔	✔		✔	49.46	50.67	39.24	57.00	65.53	62.92	52.73	50.84	47.92
4	✔	✔	✔	✔	**54.29**	**53.98**	**42.64**	**61.71**	**68.45**	**64.40**	**60.78**	**59.69**	**52.91**

Table 3. Ablation studies conducted on the SYSU-MM01 and RegDB datasets under unpaired settings. $*$ refers to $\alpha = 0.5$ in both unpaired datasets. "CFM" denotes the cross-modality feature mapping module, while "SL" and "DL" mean the static learning method and the dynamic learning method in 3.3, respectively. Rank-1 accuracy (%), mAP (%), and mINP (%) are reported.

including random cropping, random flipping and random erasing, are applied. At the beginning of each epoch, we perform the DBSCAN [7] clustering to generate pseudo labels in each modality independently. The learnable scaling and shifting parameters, γ and ζ, are initialized to 1 and 0, respectively, and the constant ϵ is set to 10^{-5}. The temperature factors τ and τ' are 0.05 and 0.6, respectively. We train the model in total of 50 epochs, in which the previous 30 epochs are used for pre-training, and CFM and SDC are executed in the last 20 epochs. All other experimental settings follow the previous work [32]. For consistency, we use the same MCL framework in paired settings to demonstrate its general applicability.

4.3. Comparison with State-of-the-Arts

To validate the efficacy of our method, we comprehensively compare it with state-of-the-art approaches under both unpaired and paired settings, as shown in Tab. 1 and Tab. 2.

Comparison with unsupervised methods under unpaired settings. The experiments demonstrate that our MCL significantly outperforms existing unsupervised methods under unpaired settings at various ratios. These considerable gains benefit from the insightful design of our method for USL-VI-ReID. There are two major advantages of our method: 1) We use a mapping-based solution to effectively address the lack of cross-modality correlations in unpaired scenarios by synthesizing fake paired data while preserving discriminative identity. 2) Our collaborative learning strategy achieves a dual-level alignment paradigm for cluster- and instance-level optimization.

Comparison with supervised methods under paired settings. We compare our method with several recent supervised visible-infrared ReID approaches under paired settings. Notably, although supervised methods rely on accurate manual annotations, our method achieves competitive performance of certain baselines (*e.g.* DFLN-ViT [48]).

Comparison with unsupervised methods under paired settings. We also evaluate our method against advanced USL-VI-ReID approaches under paired settings. Among these, H2H [16], OTLA [24] and DOTLA [4] need an extra annotated visible dataset. In Tab. 2, our method is significantly better than all existing USL-VI-ReID methods, demonstrating the effectiveness under paired settings.

4.4. Ablation Study

The performance boost of the MCL framework in the task of USL-VI-ReID under unpaired settings mainly comes from the proposed CFM module and the collaborative learning strategy with Static Learning (SL) and Dynamic Learning (DL). We validate the effectiveness of each component by conducting ablation studies on the SYSU-MM01 and RegDB under unpaired settings where α is representatively set to 0.5, which incorporate a balance of paired and unpaired scenarios. Results are shown in Tab. 3.

Baseline in index 1 denotes that we directly train the model on the unpaired SYSU-MM01 and RegDB tasks with ADCA [32] method, using the feature extractor from TransReID [10] as the backbone. Although ADCA has a promising performance under unpaired settings, it is observed that the baseline only achieves 47.29% mAP on SYSU-MM01 (all search) and 48.6% mAP on RegDB (visible to infrared). Therefore, directly using ADCA method can hardly tackle the problem of unpaired settings for the USL-VI-ReID task.

Effectiveness of SL. Index 2 means the static learning strategy with CFM module. Compared with baseline, SL improves the performance of 5.19% and 5.92% mAP on the unpaired SYSU-MM01 (all search) and RegDB (visible to infrared). The main gain is achieved by the design of the multi-memory banks in both the original and pseudo-identity space, which additionally cluster all mapped features for contrastive learning, allowing the model to capture certain modality-invariant features in scenarios with cross-modal identity mismatches.

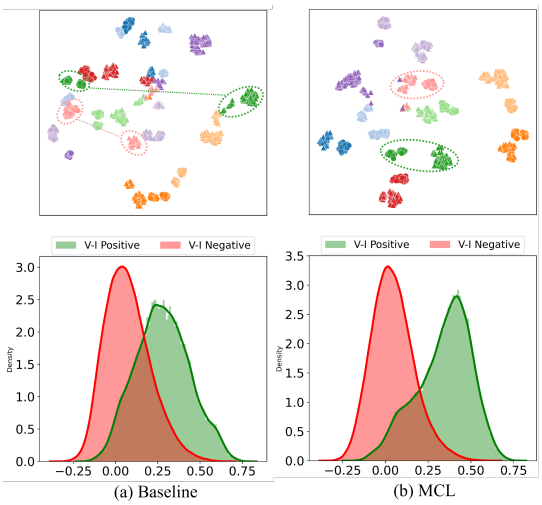
Effectiveness of DL. Index 3 represents the dynamic learning strategy with CFM module. Compared with baseline, the significant improvements demonstrate the effectiveness of DL. The improvements are 3.38% and 2.24% mAP on the unpaired SYSU-MM01 (all search) and RegDB (visible to infrared). DL can further use fake cross-modality features and pull them towards their corresponding actual samples, facilitating instance-level cross-modality learning.

Effectiveness of Collaborative Learning. Index 4 denotes the collaborative learning strategy with SL and DL. In comparison, the collaborative learning brings consistent improvement in all settings, which shows the effectiveness of static-dynamic collaborative learning in mitigating intra-modality discrepancies while enhancing cross-modality correspondence alignment.

4.5. Further Analysis

Hyper-parameter Analysis for η. We explore the influence of hyper-parameter η in Eq. 19, as presented in the **supplementary materials**. When $\eta = 1.0$, the method achieves a balance in static and dynamic learning.

Visualization. In Fig. 3, it shows the t-SNE [22] visualization of 20 randomly selected identities from the unpaired SYSU-MM01 dataset with $\alpha = 0.5$. Compared to the baseline, MCL brings infrared and visible positive samples closer together and effectively increases the separation from negative pairs, thus improving performance under unpaired settings. However, some same-identity samples remain dispersed, indicating challenges for unpaired USL-VI-ReID.



(a) Baseline (b) MCL

Figure 3. The t-SNE (top) and similarity distribution (bottom) visualizations for randomly sampled identities. Colors represent identities; circles and triangles denote visible and infrared samples, respectively. The similarity distribution represents similarity scores of cross-modality positive and negative pairs, where larger separation indicates better identity discrimination.

5. CONCLUSION

This paper addresses unsupervised visible-infrared person re-identification (USL-VI-ReID) under unpaired settings, which is a prevalent yet underexplored challenge in real-world surveillance systems. While existing USL-VI-ReID methods often overlook the critical issue of aligning cross-modality relations without paired data, we propose the Mapping and Collaborative Learning (MCL) framework to bridge this gap. We integrates a cross-modality feature mapping module to establish inter-modal correlations and a static-dynamic collaborative learning strategy to refine discriminative representations in unpaired scenarios. Extensive experiments on SYSU-MM01 and RegDB demonstrate that MCL not only outperforms state-of-the-art unsupervised methods under unpaired settings but also achieves competitive performance in paired scenarios, pushing USL-VI-ReID to real-world deployment.

Limitations and Future Work. While effective, MCL still faces challenges in extremely unpaired settings and requires relatively high training cost. Future work will focus on improving robustness in such scenarios and designing lightweight variants for efficient deployment.

Acknowledgments

This work is partially supported by National Natural Science Foundation of China under Grants (62176188, 62225113, 623B2080), the Innovative Research Group Project of Hubei Province under Grants (2024AFA017), the Major Project of Science and Technology Innovation of Hubei Province (2024BCA003), Postdoctoral Fellowship Program of China Postdoctoral Science Foundation (GZC20241268, 2024M762479), Hubei Postdoctoral Talent Introduction Program (2024HBBHJD070) and Hubei Provincial Natural Science Foundation of China (2025AFB219).The numerical calculations in this paper had been supported by the super-computing system in the Supercomputing Center of Wuhan University.

References

[1] Ying-Cong Chen, Xiatian Zhu, Wei-Shi Zheng, and Jian-Huang Lai. Person re-identification by camera correlation aware feature augmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(2):392–408, 2017. 1

[2] Zhong Chen, Zhizhong Zhang, Xin Tan, Yanyun Qu, and Yuan Xie. Unveiling the power of clip in unsupervised visible-infrared person re-identification. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3667–3675, 2023. 2, 7

[3] De Cheng, Lingfeng He, Nannan Wang, Shizhou Zhang, Zhen Wang, and Xinbo Gao. Efficient bilateral cross-modality cluster matching for unsupervised visible-infrared person reid. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1325–1333, 2023. 7

[4] De Cheng, Xiaojian Huang, Nannan Wang, Lingfeng He, Zhihui Li, and Xinbo Gao. Unsupervised visible-infrared person reid by collaborative learning with neighbor-guided label refinement. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7085–7093, 2023. 2, 7

[5] Zuozhuo Dai, Guangyuan Wang, Weihao Yuan, Siyu Zhu, and Ping Tan. Cluster contrast for unsupervised person re-identification. In *Proceedings of the Asian Conference on Computer Vision*, pages 1142–1160, 2022. 5

[6] Yunhao Du, Zhicheng Zhao, and Fei Su. Yyds: Visible-infrared person re-identification with coarse descriptions. *arXiv preprint arXiv:2403.04183*, 2024. 7

[7] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, pages 226–231, 1996. 7

[8] Xingye Fang, Yang Yang, and Ying Fu. Visible-infrared person re-identification via semantic alignment and affinity inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11270–11279, 2023. 7

[9] Wenhang Ge, Chunyan Pan, Ancong Wu, Hongwei Zheng, and Wei-Shi Zheng. Cross-camera feature prediction for intra-camera supervised person re-identification across distant scenes. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3644–3653, 2021. 2

[10] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15013–15022, 2021. 6, 8

[11] Zhangyi Hu, Bin Yang, and Mang Ye. Empowering visible-infrared person re-identification with large foundation models. In *Advances in Neural Information Processing Systems*, pages 117363–117387. Curran Associates, Inc., 2024. 7

[12] Minsu Kim, Seungryong Kim, Jungin Park, Seongheon Park, and Kwanghoon Sohn. 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*, pages 18621–18632, 2023. 7

[13] Xiangyuan Lan, Shengping Zhang, Pong C Yuen, and Rama Chellappa. Learning common and feature-specific patterns: a novel multiple-sparse-representation-based tracker. *IEEE Transactions on Image Processing*, 27(4):2022–2037, 2017. 1

[14] Shu Li, Huaiyuan Wang, and Ruimin Hu. A review of public social behavior understanding and hidden groups discovery. *Journal of Image and Graphics*, 30(6):2275–2303, 2025. 2

[15] Zhiyong Li, Haojie Liu, Xiantao Peng, and Wei Jiang. Inter-intra modality knowledge learning and clustering noise alleviation for unsupervised visible-infrared person re-identification. *IEEE Transactions on Knowledge and Data Engineering*, 2024. 1, 7

[16] Wenqi Liang, Guangcong Wang, Jianhuang Lai, and Xiaohua Xie. Homogeneous-to-heterogeneous: Unsupervised learning for rgb-infrared person re-identification. *IEEE Transactions on Image Processing*, 30:6392–6407, 2021. 2, 7

[17] Xinyu Lin, Jinxing Li, Zeyu Ma, Huafeng Li, Shuang Li, Kaixiong Xu, Guangming Lu, and David Zhang. Learning modal-invariant and temporal-memory for video-based visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20973–20982, 2022. 2

[18] Jialun Liu, Yifan Sun, Feng Zhu, Hongbin Pei, Yi Yang, and Wenhui Li. Learning memory-augmented unidirectional metrics for cross-modality person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19366–19375, 2022. 2

[19] Yongheng Qian and Su-Kit Tang. Multi-scale contrastive learning with hierarchical knowledge synergy for visible-infrared person re-identification. *Sensors (Basel, Switzerland)*, 25(1):192, 2025. 1

[20] Jiangming Shi, Xiangbo Yin, Yeyun Chen, Yachao Zhang, Zhizhong Zhang, Yuan Xie, and Yanyun Qu. Multi-memory matching for unsupervised visible-infrared person re-identification. In *European Conference on Computer Vision*, pages 456–474. Springer, 2024. 1, 6, 7

[21] Xiao Teng, Long Lan, Dingyao Chen, Kele Xu, and Nan Yin. Relieving universal label noise for unsupervised visible-

infrared person re-identification by inferring from neighbors. *arXiv preprint arXiv:2412.12220*, 2024. 6, 7

[22] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9 (11), 2008. 8

[23] Dongkai Wang and Shiliang Zhang. Unsupervised person re-identification via multi-label classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 2

[24] Jiangming Wang, Zhizhong Zhang, Mingang Chen, Yi Zhang, Cong Wang, Bin Sheng, Yanyun Qu, and Yuan Xie. Optimal transport for label-efficient visible-infrared person re-identification. In *European Conference on Computer Vision*, pages 93–109. Springer, 2022. 2, 7

[25] Yuhao Wang, Xuehu Liu, Pingping Zhang, Hu Lu, Zhengzheng Tu, and Huchuan Lu. Top-reid: Multi-spectral object re-identification with token permutation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5758–5766, 2024. 2

[26] Zheng Wang, Zhixiang Wang, Yinqiang Zheng, Yang Wu, Wenjun Zeng, and Shin'ichi Satoh. Beyond intra-modality: A survey of heterogeneous person re-identification. *arXiv preprint arXiv:1905.10048*, 2019. 1

[27] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 79–88, 2018. 6

[28] Ziyu Wei, Xi Yang, Nannan Wang, and Xinbo Gao. Syncretic modality collaborative learning for visible infrared person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 225–234, 2021. 2

[29] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. Rgb-infrared cross-modality person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5380–5389, 2017. 2, 6

[30] Ancong Wu, Wei-Shi Zheng, Shaogang Gong, and Jianhuang Lai. Rgb-ir person re-identification by cross-modality similarity preservation. *International Journal of Computer Vision*, 128(6):1765–1785, 2020. 1

[31] Zesen Wu and Mang Ye. 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*, pages 9548–9558, 2023. 1, 2, 7

[32] Bin Yang, Mang Ye, Jun Chen, and Zesen Wu. Augmented dual-contrastive aggregation learning for unsupervised visible-infrared person re-identification. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2843–2851, 2022. 1, 2, 3, 6, 7, 8

[33] Bin Yang, Jun Chen, and Mang Ye. Towards grand unified representation learning for unsupervised visible-infrared person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11069–11079, 2023. 2, 7

[34] Fengxiang Yang, Zhun Zhong, Zhiming Luo, Yuanzheng Cai, Yaojin Lin, Shaozi Li, and Nicu Sebe. Joint noise-tolerant learning and meta camera shift adaptation for unsupervised person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4855–4864, 2021. 2

[35] Mouxing Yang, Zhenyu Huang, Peng Hu, Taihao Li, Jiancheng Lv, and Xi Peng. Learning with twin noisy labels for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14308–14317, 2022. 2

[36] Yang Yang, Tianzhu Zhang, Jian Cheng, Zengguang Hou, Prayag Tiwari, Hari Mohan Pandey, et al. Cross-modality paired-images generation and augmentation for rgb-infrared person re-identification. *Neural Networks*, 128:294–304, 2020. 2

[37] Yiming Yang, Weipeng Hu, and Haifeng Hu. Progressive cross-modal association learning for unsupervised visible-infrared person re-identification. *IEEE Transactions on Information Forensics and Security*, 2025. 2, 6, 7

[38] Mang Ye, Xiangyuan Lan, Qingming Leng, and Jianbing Shen. Cross-modality person re-identification via modality-aware collaborative ensemble learning. *IEEE Transactions on Image Processing*, 29:9387–9399, 2020. 1

[39] Mang Ye, Jianbing Shen, David J. Crandall, Ling Shao, and Jiebo Luo. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In *European Conference on Computer Vision*, pages 229–247. Springer, 2020. 6

[40] Mang Ye, Weijian Ruan, Bo Du, and Mike Zheng Shou. Channel augmented joint learning for visible-infrared recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13567–13576, 2021. 2, 6

[41] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):2872–2893, 2021. 1, 6, 7

[42] Mang Ye, Shuoyi Chen, Chenyue Li, Wei-Shi Zheng, David Crandall, and Bo Du. Transformer for object re-identification: A survey. *International Journal of Computer Vision*, pages 1–31, 2024. 1

[43] Mang Ye, Zesen Wu, and Bo Du. Dual-level matching with outlier filtering for unsupervised visible-infrared person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3815–3829, 2025. 2

[44] Hao Yu, Xu Cheng, Wei Peng, Weihao Liu, and Guoying Zhao. Modality unifying network for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11185–11195, 2023. 7

[45] Xinyu Zhang, Dongdong Li, Zhigang Wang, Jian Wang, Errui Ding, Javen Qinfeng Shi, Zhaoxiang Zhang, and Jingdong Wang. Implicit sample extension for unsupervised person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7369–7378, 2022. 5

[46] Yukang Zhang and Hanzi Wang. Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2153–2162, 2023. 6

[47] Zhizhong Zhang, Jiangming Wang, Xin Tan, Yanyun Qu, Junping Wang, Yong Xie, and Yuan Xie. Mutual information guided optimal transport for unsupervised visible-infrared person re-identification. *arXiv preprint arXiv:2407.12758*, 2024. 6

[48] Jiaqi Zhao, Hanzheng Wang, Yong Zhou, Rui Yao, Silin Chen, and Abdulmotaleb El Saddik. Spatial-channel enhanced transformer for visible-infrared person re-identification. *IEEE Transactions on Multimedia*, 25:3668–3680, 2022. 7

[49] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1116–1124, 2015. 6

[50] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016. 1

[51] Yi Zheng, Shixiang Tang, Guolong Teng, Yixiao Ge, Kaijian Liu, Jing Qin, Donglian Qi, and Dapeng Chen. Online pseudo label generation by hierarchical cluster dynamics for adaptive person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8371–8381, 2021. 2