Shallow-Deep Collaborative Learning for Unsupervised Visible-Infrared Person Re-Identification

Bin Yang Jun Chen* Mang Ye *
National Engineering Research Center for Multimedia Software,
School of Computer Science, Hubei Luojia Laboratory, Wuhan University, Wuhan, China
https://github.com/yangbincv/SDCL

Abstract

Unsupervised visible-infrared person re-identification (US-VI-ReID) centers on learning a cross-modality retrieval model without labels, reducing the reliance on expensive cross-modality manual annotation. Previous US-VI-ReID works gravitate toward learning cross-modality information with the deep features extracted from the ultimate layer. Nevertheless, interfered by the multiple discrepancies, solely relying on deep features is insufficient for accurately learning modality-invariant features, resulting in negative optimization. The shallow feature from the shallow layers contains nuanced detail information, which is critical for effective cross-modality learning but is disregarded regrettably by the existing methods. To address the above issues, we design a Shallow-Deep Collaborative Learning (SDCL) framework based on the transformer with shallow-deep contrastive learning, incorporating Collaborative Neighbor Learning (CNL) and Collaborative Ranking Association (CRA) module. Specifically, CNL unveils the intrinsic homogeneous and heterogeneous collaboration which are harnessed for neighbor alignment, enhancing the robustness in a dynamic manner. Furthermore, CRA associates the cross-modality labels with the ranking association between shallow and deep features, furnishing valuable supervision for cross-modality learning. Extensive experiments validate the superiority of our method, even outperforming certain supervised counterparts.

1. Introduction

Person re-identification (ReID) aims at matching the same person image captured by non-overlapping cameras. In recent times, ReID has garnered significant attention from the computer vision research community, owing to its pivotal role in the context of intelligent video surveillance applications [2, 18, 19, 23, 25, 28–30, 39, 41, 44, 46, 57, 59, 62, 63]. The rank-1 accuracy has exhibited encouraging outcomes in supervised person ReID. Notably, these person images are typically acquired by visible cameras within well-illuminated environments. However, it is crucial to recognize that visible cameras often fail to capture sufficient information of individuals in low-light settings, consequently constraining the practicality of single-modality ReID in the context of night-time surveillance [1, 7, 15, 21, 40, 42, 45, 53, 61].

In contrast, recently deployed cameras possess the capability to seamlessly transition into far/near-infrared mode during nighttime operations [50, 55]. Various techniques [12, 20, 22, 27, 34, 47, 52, 54–58, 62] have been proposed for visible-infrared person re-identification (VI-ReID), resulting in commendable accuracy levels. Notably, these achievements are facilitated by extensive human-labeled cross-modality datasets. However, annotating cross-modality datasets demands more resources compared to the annotation of single-modality ReID datasets. This

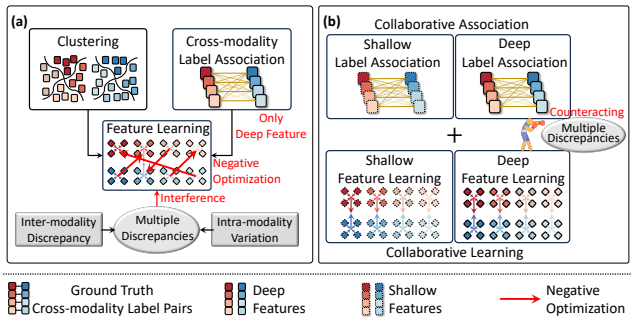


Figure 1. Illustration of the motivation. Previous advanced works for US-VI-ReID mainly focus on learning features and associating cross-modality labels with deep features. However, interfered by the multiple discrepancies, solely relying on deep features is insufficient for accurately learning modality-invariant features, resulting in negative optimization, as shown in (a). In this paper, we incorporate shallow and deep features with collaborative learning and label association to remedy these problems, as shown in (b).

*Corresponding Author.

intricate annotation task presents a formidable and pivotal challenge, ultimately giving rise to the crucial task of unsupervised VI-ReID (US-VI-ReID) for reducing the cost of expensive cross-modality annotations.

Some prior research endeavors [4, 5, 24, 33, 37, 48, 50, 51] have proposed preliminary solutions for US-VI-ReID. These solutions gravitate toward learning cross-modality feature with the deep features extracted from the ultimate layer. Nevertheless, interfered by the multiple discrepancies, *i.e*., intra-modality variation and inter-modality discrepancy, solely relying on deep features is insufficient for accurately learning modality-invariant features, resulting in negative optimization, as shown in Fig. 1. It is important to recognize that the shallow features originating from the modality-specific shallow layers harbor a wealth of nuanced detail information pertaining to pedestrian attributes, which is critical for unsupervised cross-modality representation learning but is disregarded regrettably by the existing methods. To harness the complete potential of the shallow information, we devise a comprehensive Shallow-Deep Collaborative Learning (SDCL) framework based on the transformer with shallow-deep contrastive learning, incorporating Collaborative Neighbor Learning (CNL) and Collaborative Ranking Association (CRA) module. Specifically, CNL unveils the intrinsic homogeneous and heterogeneous collaboration as the constraints for seeking reliable intra-modality and inter-modality neighbor learning, thereby guaranteeing the cultivation of modality-invariant and discriminative representations in a dynamic manner. Furthermore, CRA is developed with a global perspective to exploit cross-modality ranking consistency between deep and shallow features, associating the label of two modalities in a collaborative manner and furnishing valuable cross-modality supervision for cross-modality learning. Ultimately, within the shallow-deep collaborative learning framework with CNL and CRA modules, we acquire a robust representation, effectively mitigating the cross-modality discrepancy under unsupervised conditions.

The main contributions can be summarized as follows:

- We propose a shallow-deep collaborative learning framework based on the transformer architecture. This framework facilitates the learning of robust representation, effectively countering the cross-modality discrepancy through the collaboration of shallow and deep features.
- We propose a collaborative neighbor learning module to formulate dependable intra-modality and cross-modality neighbor learning, enabling the model to capture modality-invariant and discriminative features.
- We propose a collaborative ranking association module to explore intra-modality and cross-modality ranking consistencies, unifying the cross-modality labels and providing invaluable cross-modality supervision.
- Extensive experiments validate that our SDCL framework surpasses existing methods on two mainstream VI-ReID benchmarks, consistently improving the unsupervised cross-modality retrieval performance.

2. Related Work

2.1. Supervised Visible-Infrared Person ReID

Currently, advanced supervised VI-ReID focuses on bridging the gap between the two modalities and learning robust representations against modality discrepancy. Ye *et al*. [55] developed a dynamic tri-level relation mining (DTRM) framework simultaneously to explore the cross-modality relation cues of channel-level, part-level intra-modality, and graph-level. CAJ [56] proposed a channel-mixed learning to handle the intra-modality and cross-modality variations by randomly exchanging the color channels. SGIEL [11] proposed a shape-erased feature learning paradigm, jointly learning shape-related feature in one subspace and shape-erased features in the orthogonal complement.

Nevertheless, the remarkable performances exhibited by these methods require extensive human-labeled cross-modality datasets. In this work, we pivot our focus to the realm of unsupervised visible-infrared person ReID, where the luxury of identity annotations is absent, presenting important applications for real-world VI-ReID deployments.

2.2. Unsupervised Single-Modality Person ReID

Unsupervised single-modality ReID endeavors to train the ReID model using unlabeled data captured by visible cameras. The majority of studies have embraced cluster algorithms to derive pseudo labels for optimizing the model. Cluster Contrast [9] utilized a distinctive cluster representation to delineate each cluster, addressing the issue of cluster inconsistency. Chen *et al*. [3] introduced Inter-instance Contrastive Encoding (ICE), which harnesses inter-instance pairwise similarity scores to enhance preceding class-level contrastive ReID methodologies. IICS [49] tackled the unsupervised ReID challenge by decomposing the similarity computation into two stages, namely, the intra-domain and inter-domain computations, respectively. Dai *et al*. [8] introduced a dual-refinement approach to concurrently enhance pseudo labels during the offline clustering phase and refine features during the online training phase, augmenting label purity and feature discriminability.

The above methods are dedicated to addressing issues in single-modality ReID. When employed in the context of US-VI-ReID, these approaches encounter challenges stemming from cross-modality discrepancies and the absence of cross-modality (visible-infrared) identity labels. This impedes the learning of inter-modality feature and the generation of reliable cross-modality pseudo labels.

2.3. Unsupervised Visible-Infrared Person ReID

Several studies [24, 37, 48, 50] represent initial endeavors in the realm of US-VI-ReID. Yang *et al*. [50] introduced a novel Augmented Dual-Contrastive Aggregation (ADCA) learning framework founded on the principles of homogeneous joint learning and heterogeneous aggregation, establishing a robust baseline for purely unsupervised VI-ReID. PGM [48] developed a progressive graph matching method to systematically extract cross-modality correspondences, formulating correspondence mining as a graph-matching process. H2H [24] introduced a homogeneous-to-heterogeneous approach through two-stage learning, incorporating Market1501 [64] as additional labeled data. CHCR [33] proposed a cross-modality hierarchical clustering and refinement method by promoting modality-invariant feature learning and improving the reliability of pseudo-labels.

Nevertheless, the above methods primarily concentrate on feature learning with deep features. However, interfered by the intra-modality variation and inter-modality discrepancy, solely relying on deep features is insufficient for accurately learning modality-invariant features, resulting in negative optimization. Our approach integrates shallow and deep features through collaborative feature learning and cross-modality label association, which is distinguished from previous works, surpassing the performance of existing US-VI-ReID methods.

3. Proposed Method

The framework of shallow-deep collaborative learning (SDCL) is shown in Fig. 2, incorporating the Collaborative Neighbor Learning (CNL) and Collaborative Ranking Association (CRA) module. SDCL adopts a dual-path transformer architecture with dual-contrastive learning [50], which has two shallow modality-specific patch embedding layers and a modality-shared transformer. Instance memory and cluster memory are constructed for shallow embeddings and deep features within each modality. Instance memory stores all training image features. Cluster memory is built by averaging the instance features with the same pseudo labels. With the above memories, SDCL develops the CNL module to exploit homogeneous and heterogeneous collaboration as the constraints for seeking reliable intra-modality and inter-modality neighbor learning. CRA explores the ranking consistency of cross-modality shallow embeddings and deep features to associate reliable cross-modality identities. With the collaboration of shallow embeddings and deep features within CNL and CRA modules, SDCL learns better representation for cross-modality retrieval.

3.1. Shallow-deep Contrastive Learning

We first introduce shallow-deep contrastive learning based on augmented dual-contrast learning [50] with a dual-path transformer architecture.

Cluster Memory Initialization. At the beginning of each training epoch, we construct shallow and deep cluster memories for each modality by averaging the shallow and deep feature of one cluster, which can be denoted as:

$$\phi_k^{vs} = \frac{1}{|\mathcal{H}_k^v|} \sum_{u_n^{vs} \in y_k^v} u_n^{vs}, \tag{1}$$

$$\phi_k^{vd} = \frac{1}{|\mathcal{H}_k^v|} \sum_{u_n^{vd} \in y_k^v} u_n^{vd}, \tag{2}$$

$$\phi_l^{rs} = \frac{1}{|\mathcal{H}_l^r|} \sum_{u_n^{rs} \in y_l^r} u_n^{rs}, \tag{3}$$

$$\phi_l^{rd} = \frac{1}{|\mathcal{H}_l^r|} \sum_{u_n^{rd} \in y_l^r} u_n^{rd}, \tag{4}$$

where u_n^{r*} and u_n^{v*} are infrared and visible instance features, respectively. u_n^{*s} and u_n^{*d} are shallow and deep instance features, respectively. $\mathcal{H}_k$ indicates the k-th cluster set and $|\cdot|$ counts the number of samples of a set. Meanwhile, we store the instance features in shallow and deep infrared and visible instance memories U^{rs}, U^{rd}, U^{vs} and U^{vd}.

Cluster Memory Updating. During each training iteration, the deep and shallow cluster memories of two modalities by a momentum updating strategy:

$$\phi_k^{(\delta)} \leftarrow \alpha \phi_k^{(\delta-1)} + (1-\alpha) q, q \in y_k, \tag{5}$$

where q is the query features sampled from training set in each training iteration, respectively. α is the momentum factor. δ is the training iteration number.

Contrastive Loss. Given shallow and deep visible and infrared query q^{vs}, q^{vd}, q^{rs}, and q^{rs}, we compute the contrastive loss for visible modality by the following equations:

$$\mathcal{L}_{id}^{vs} = -\log \frac{\exp\left(q^{vs} \cdot \phi_+^{vs}/\tau\right)}{\sum_{k=0}^{K} \exp\left(q^{vs} \cdot \phi_k^{vs}/\tau\right)}, \tag{6}$$

$$\mathcal{L}_{id}^{vd} = -\log \frac{\exp\left(q^{vd} \cdot \phi_+^{vd}/\tau\right)}{\sum_{k=0}^{K} \exp\left(q^{vd} \cdot \phi_k^{vd}/\tau\right)}, \tag{7}$$

where ϕ_+ is the positive memory corresponding to the pseudo label of q and the τ is a temperature. The infrared contrastive loss $\mathcal{L}_{id}^{rs}$ and $\mathcal{L}_{id}^{rd}$ are obtained by the same way.

We optimize the modality-specific shallow embedding layers and deep modality-shared layers by combining deep and shallow contrastive loss:

$$\mathcal{L}_{id} = \mathcal{L}_{id}^{vd} + \mathcal{L}_{id}^{rd} + \mathcal{L}_{id}^{vs} + \mathcal{L}_{id}^{rs} \tag{8}$$

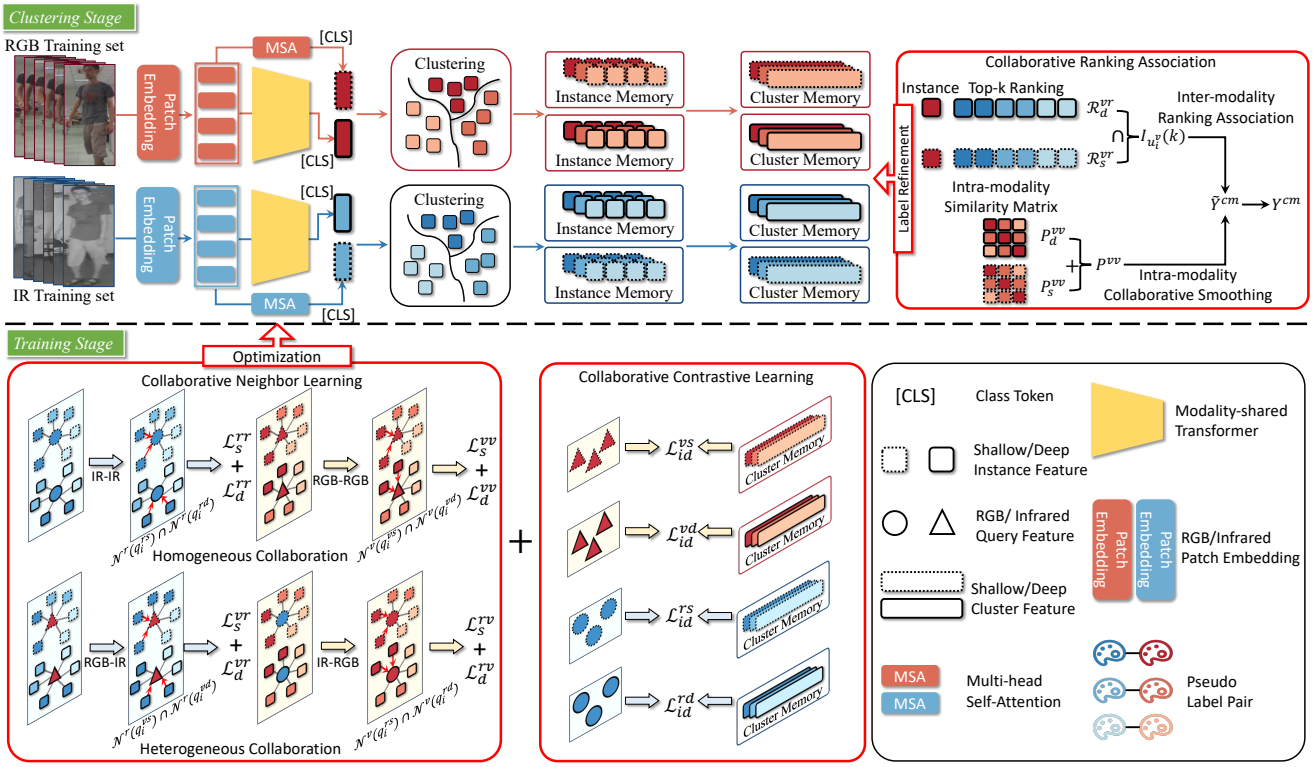


Figure 2. Illustration of shallow-deep collaborative learning. It comprises shallow-deep contrastive learning, collaborative neighbor learning, and collaborative ranking association. SDCL finds reliable neighborhood and cross-modality labels with shallow-deep collaborations, enhancing the robustness of learned representation.

3.2. Collaborative Neighbor Learning

Our basic rationale for the CNL module is that deep and shallow features are complementary, and their intrinsic consistency relationship can be used to constrain intra-modality and inter-modality optimization, formulating collaborative learning to enhance the robustness against intra-modality and inter-modality variations. From the above perspective, CNL explores essential homogeneous and heterogeneous shallow-deep collaborations to seek reliable intra-modality and inter-modality neighbor learning.

Homogeneous and Heterogeneous Collaboration. Given a query q, we can get the similarity $s(q_i, u_j)$ between the query and each instance in the training set by:

$$s(q_i, u_j) = \frac{q_i \cdot u_j}{||q_i||_2 ||u_j||_2}, \tag{9}$$

where q_i and u_j come from shallow or deep features of visible or infrared modality to get multiple types of homogeneous and heterogeneous shallow or deep query-instance similarity $s(q_i^{vs}, u_j^{vs})$, $s(q_i^{rs}, u_j^{rs})$, $s(q_i^{vs}, u_j^{rs})$, $s(q_i^{rs}, u_j^{vs})$, $s(q_i^{vd}, u_j^{vd})$, $s(q_i^{rd}, u_j^{rd})$, $s(q_i^{vd}, u_j^{rd})$, and $s(q_i^{rd}, u_j^{vd})$.

With the above multiple types of similarities, we seek reliable shallow and deep intra-modality and inter-modality neighbors, formulating collaborative neighbor learning. The intra-modality neighbors can be defined as follows:

$$\mathcal{N}^v(q_i^v) = \{\mathcal{N}^v(q_i^{vs}) \cap \mathcal{N}^v(q_i^{vd})\}, \tag{10}$$

$$\mathcal{N}^r(q_i^r) = \{\mathcal{N}^r(q_i^{rs}) \cap \mathcal{N}^r(q_i^{rd})\}, \tag{11}$$

where $\mathcal{N}^v(q_i^{v*})$ and $\mathcal{N}^r(q_i^{r*})$ are neighborhood sets searched by visible-visible and infrared-infrared similarity, respectively. $* \in \{s, d\}$ represents the shallow and deep feature. The $\mathcal{N}^v(q_i^{vs})$ and $\mathcal{N}^r(q_i^{vd})$ can be obtained by:

$$\mathcal{N}^v(q_i^{vs}) = \{u_j^{vs} | s(q_i^{vs}, u_j^{vs}) > \gamma \cdot \max_{j=1\ldots N^v} s(q_i^{vs}, u_j^{vs})\}, \tag{12}$$

$$\mathcal{N}^v(q_i^{vd}) = \{u_j^{vd} | s(q_i^{vd}, u_j^{vd}) > \gamma \cdot \max_{j=1\ldots N^v} s(q_i^{vd}, u_j^{vd})\}, \tag{13}$$

where the N^v denotes the number of visible instances and the γ is the positive neighbor selection range. The $\mathcal{N}^r(q_i^{rs})$ and $\mathcal{N}^r(q_i^{rd})$ can be calculated by similar manner.

Given query q_i^v or q_i^r, the inter-modality neighbors can be denoted as follows:

$$\mathcal{N}^r(q_i^v) = \{\mathcal{N}^r(q_i^{vs}) \cap \mathcal{N}^r(q_i^{vd})\}, \tag{14}$$

$$\mathcal{N}^v(q_i^r) = \{\mathcal{N}^v(q_i^{rs}) \cap \mathcal{N}^v(q_i^{rd})\}, \tag{15}$$

where $\mathcal{N}^r(q_i^{vs})$ and $\mathcal{N}^r(q_i^{vd})$ are defined as follows:

$$\mathcal{N}^r(q_i^{vs}) = \{u_j^{rs} | s(q_i^{vs}, u_j^{rs}) > \gamma \cdot \max_{j=1...N^r} s(q_i^{vs}, u_j^{rs})\}, \tag{16}$$

$$\mathcal{N}^r(q_i^{vd}) = \{u_j^{rd} | s(q_i^{vd}, u_j^{rd}) > \gamma \cdot \max_{j=1...N^r} s(q_i^{vd}, u_j^{rd})\}, \tag{17}$$

where the N^r denotes the number of infrared instances. Similarly, we can get the $\mathcal{N}^v(q_i^{rs})$ and $\mathcal{N}^v(q_i^{rd})$. With the above procedure, we get the $\mathcal{N}^v(q_i^v)$, $\mathcal{N}^r(q_i^r)$, $\mathcal{N}^r(q_i^v)$ and $\mathcal{N}^v(q_i^r)$, which are the reliable neighborhood set searched by the constraints of homogeneous or heterogeneous shallow-deep collaborations.

Collaborative Neighbor Learning. With the neighborhood set $\mathcal{N}^v(q_i^v)$, $\mathcal{N}^r(q_i^r)$, $\mathcal{N}^r(q_i^v)$ and $\mathcal{N}^v(q_i^r)$, we can perform neighbor learning. Given query q_i^{vs} and q_i^{vs}, we can obtain the expression of the visible-visible shallow and deep neighbor learning by:

$$\mathcal{L}_s^{vv} = -\frac{1}{N^b} \sum_{i=1}^{N^b} \sum_{j \in \mathcal{N}^v(q_i^v)} \log \frac{\exp\left(s(q_i^{vs}, u_j^{vs})/\tau\right)}{\sum_{n=1}^{N^v} \exp\left(s(q_i^{vs}, u_n^{vs})/\tau\right)}, \tag{18}$$

$$\mathcal{L}_d^{vv} = -\frac{1}{N^b} \sum_{i=1}^{N^b} \sum_{j \in \mathcal{N}^v(q_i^v)} \log \frac{\exp\left(s(q_i^{vd}, u_j^{vd})/\tau\right)}{\sum_{n=1}^{N^v} \exp\left(s(q_i^{vd}, u_n^{vd})/\tau\right)}, \tag{19}$$

$$\mathcal{L}^{vv} = \mathcal{L}_d^{vv} + \lambda_s \mathcal{L}_s^{vv}, \tag{20}$$

where N^b is the batch size of query q_i. Similarly, the neighbor learning for infrared-infrared $\mathcal{L}^{rr}$, infrared-visible $\mathcal{L}^{rv}$, and visible-infrared $\mathcal{L}^{vr}$ can be obtained by similar ways. The final optimization for neighbor learning is denoted by the following combination:

$$\mathcal{L}_{neighbor} = \mathcal{L}^{vv} + \mathcal{L}^{rr} + \mathcal{L}^{rv} + \mathcal{L}^{vr}. \tag{21}$$

The overall loss for SDCL is denoted as:

$$\mathcal{L}_{total} = \mathcal{L}_{neighbor} + \mathcal{L}_{id} \tag{22}$$

Discussion. In contrast to neighborhood learning in [30], which exclusively relies on deep features for the selection of positive neighbors. Under the multiple discrepancies, the similarity of individual deep features is unreliable, resulting in wrong neighbor selection and negative optimization. Our method embraces collaborative neighbor learning, which dynamically refines cross-modality supervision through shallow and deep collaboration, achieving effective cross-modality feature alignment.

3.3. Collaborative Ranking Association

The CRA module is underpinned by dual ***rationales***: **(1)** Shallow and deep features across the two modalities should have the same ranking consistency. **(2)** The ranking consistencies of shallow and deep features within one modality are more accurate. Accordingly, the CRA module contains two processes, *i.e.*, inter-modality ranking association and intra-modality ranking smoothing.

Inter-modality Ranking Association. With the similarities $s(u_i^{vs}, u_j^{rs})$ and $s(u_i^{vd}, u_j^{rd})$, we can get two visible-infrared ranking lists $\{s(u_i^{vs}, u_j^{rs}), j \in [1, N^r]\}$ and $\{s(u_i^{vd}, u_j^{rd}), j \in [1, N^r]\}$ for u_i^{vs} and u_i^{vd}, which are denoted as $\mathcal{R}_s^{vr}(u_i^{vs})$ and $\mathcal{R}_d^{vr}(u_i^{vd})$. The label of k-th similar infrared instance in two ranking lists can be represented as:

$$\tilde{y}_{u_i^{vs}}^{vr}[k] = y_{u_j^{rs}}, u_j^{rs} = \mathcal{R}_s^{vr}(u_i^{vs})[k], \tag{23}$$

$$\tilde{y}_{u_i^{vd}}^{vr}[k] = y_{u_j^{rd}}, u_j^{rd} = \mathcal{R}_d^{vr}(u_i^{vd})[k], \tag{24}$$

where $\tilde{y}_{u_i^{vs}}^{vr}[k]$ and $\tilde{y}_{u_i^{vd}}^{vr}[k]$ are refined k-th cross-modality labels of u_i^{vs} and u_i^{vd} through the ranking on visible-infrared shallow and deep similarities. Based on the ***rationale (1)***, we propose to associate the cross-modality labels with the intersection of two label sets to investigate collaborative ranking consistency:

$$I_{u_i^v}(k) = \{\tilde{y}_{u_i^{vd}}^{vr}[k] \cap \{\tilde{y}_{u_i^{vs}}^{vr}[n]\}_{n=1}^{N}\}, \tag{25}$$

where the $I_{u_i^v}(k)$ records the samples with the same identity of top-k identity in instances from top-1 to top-N. The reliable refined cross-modality label of u_i^v can be expressed by the label of maximum count in $I_{u_i^v}(k)$:

$$\tilde{y}_{u_i^v}^{cm} = \tilde{y}_{u_i^{vd}}^{vr}[k], k = \underset{k}{argmax}(\{|I_{u_i^v}(k)|, k \in [1, K]\}), \tag{26}$$

where $|\cdot|$ denotes the counting function. The $argmax$ operation traverses $I_{u_i^v}(k)$ and finds the labels with maximum number as cross-modality refined labels. Then, we convert the label list $\{\tilde{y}_{u_i^v}^{cm}, i \in [1, N^v]\}$ to the form of one-hot code matrix $\tilde{Y}^{cm} \in \mathbb{R}^{N^v \times C^r}$ by setting the column according to the refine labels to 1 and the rest to 0, where C^r is the class number of infrared modality.

Intra-modality Collaborative Smoothing. Based on the ***rationale (2)***, two shallow and deep homogeneous shallow P_s^{vv} and deep P_d^{vv} similarity matrices are constructed to investigate intra-modality ranking consistency, enhancing the precision of refined cross-modality pseudo labels by:

$$P_s^{vv}(i, j) = s(u_i^{vs}, u_j^{vs}), \tag{27}$$

$$P_d^{vv}(i, j) = s(u_i^{vd}, u_j^{vd}), \tag{28}$$

where $P_*^{vv} \in \mathbb{R}^{N^v \times N^v}$ represents the intra-modality similarity structure. In order to explore the intra-modality shallow and deep collaboration, we calculate the sum of $P_s^{vv}(i,j)$ and $P_d^{vv}(i,j)$ by:

$$P^{vv}(i,j) = P_s^{vv}(i,j) + P_d^{vv}(i,j) \tag{29}$$

where P^{vv} indicates the consistency of shallow and deep similarity matrix. We keep the 5-max values of P^{vv} in each row to 1 and the rest to 0, acquiring the ranking relations. The process of intra-modality ranking smoothing is formulated as follows:

$$Y^{cm} = P^{vv}\tilde{Y}^{cm}, \tag{30}$$

where $Y^{cm} \in \mathbb{R}^{N^v \times C^r}$ is the final refined cross-modality label matrix of visible instance. In Y^{cm}, the column number of the maximum value in each row is the refined label of samples. In our work, we refine the pseudo labels of the visible modality to the infrared modality and smooth the cross-modality labels with the shallow and deep similarity within the visible modality. Then, the infrared and refined visible labels will be used to construct a modality-shared memory for contrastive learning within two modalities [51].

Discussion. Our CRA is essentially the collaboration of shallow and deep ranking processes with a global perspective at the beginning of each training epoch, which is distinguished from previous works employing a solitary-pattern deep feature similarity [37, 48, 50, 51]. The imposition of a collaborative consistency constraint on dual-pattern similarities within CRA confers a distinct advantage, facilitating the exploration of more dependable cross-modality relations and providing valuable cross-modality supervision.

4. Experiments

4.1. Datasets and Evaluation Protocol

Datasets. We evaluate the proposed SDCL framework on two widely-used visible-infrared person ReID datasets, *i.e.*, SYSU-MM01 [45] and RegDB [32]. SYSU-MM01 dataset is collected by 6 different cameras, including 22258 visible and 11909 near-infrared images of 395 training identities. We perform ten trials of the gallery set selection [53] and calculate the average performance following existing methods with all-search and indoor-search testing mode. RegDB dataset is captured by two aligned visible and thermal camera system, which has less challenges for VI-ReID. 206 identities with 2,060 images are selected for training and this procedure is repeated 10 times following [54], and calculate the average performance with visible to thermal and thermal to visible testing mode.

Evaluation Protocols. Following existing works, Cumulative Matching Characteristics (CMC), mean Average Precision (mAP) and mean inverse negative penalty (mINP) [57] are calculated as the evaluation metrics.

4.2. Implementation Details

The proposed framework is implemented with PyTorch. SDCL incorporates the feature extractor from TransReID [17] as the backbone network with augmented dual-contrastive learning [50]. The shallow patch embedding layers are constructed with IBN and CNN module [31]. The concatenation of shallow and deep class tokens is used to calculate the cosine similarity for retrieval. DBSCAN [10] is conducted to generate pseudo labels. The visible and infrared images are resized to 288×144 before entering the network. We sample 8 pseudo identities and 16 instances for each pseudo identity for each modality within one batch. Channel augmentation [56], random crop, horizontal flipping, and random erasing are adopted for data augmentation. We adopt the SGD optimizer with the initial learning rate of $3.5e-4$ to train the model. The model is trained in a total of 50 epochs. The CRA module is added in the last 20 epochs. The λ_s in Eq. 20 is set to 0.5. The γ in Eq. 12, Eq. 13, Eq. 16, and Eq. 17 is set to 0.9. The N and K in Eq. 25 and Eq. 26 are set to 20 for SYSU-MM01 and 10 for RegDB, respectively. The contrastive learning settings follow [50].

4.3. Comparison with State-of-the-art Methods

In Table 1, our SDCL is compared with supervised and unsupervised VI-ReID methods on two benchmarks including SYSU-MM01 and RegDB.

Comparison with Unsupervised Methods. As reported in Table 1, our method exhibits superior performance compared to the current advanced unsupervised methods. More precisely, our SDCL attains a remarkable 64.49% and 86.91% rank-1 accuracy on SYSU-MM01 (all search) and RegDB (visible to infrared), respectively. Compared with the best current method GUR [51], our method also exceeds the about 1% and 10% rank-1 accuracy on SYSU-MM01 (all search) and RegDB (visible to infrared), respectively. In comparison to DPIS [35], a uni-semi-supervised method employing visible annotations for training, our SDCL still continues to maintain a leading position.

Comparison with Supervised Methods. Our comparison further includes 14 well-known supervised methods for reference. The comparisons with these methodologies unequivocally illustrate that our SDCL framework outperforms several supervised methods, including AGW [57], MSO [12], and DDAG [54], and achieves a competitive performance compared with MCLNet[16]. These considerable gains benefit from the insightful design of a shallow-deep collaborative learning framework. Our method facilitates the cultivation of modality-invariant features through the synergistic collaboration of shallow and deep features. Guided by our innovative solutions, SDCL surpasses prevailing unsupervised methods.

	Methods	Venue	SYSU-MM01						RegDB					
			All Search			Indoor Search			Visible to Infrared			Infrared to Visible		
			$r1$	mAP	mINP	$r1$	mAP	mINP	$r1$	mAP	mINP	$r1$	mAP	mINP
Supervised	DDAG [54]	ECCV-20	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 [57]	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
	CA [56]	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 [47]	CVPR-21	70.58	68.24	-	76.74	80.95	-	82.8	80.7	-	83.7	80.9	-
	MSO [12]	MM-21	58.70	56.42	-	63.09	70.31	-	73.6	66.9	-	74.6	67.5	-
	AGM [26]	MM-21	69.63	66.11	52.24	74.68	78.30	74.00	88.40	81.45	68.51	85.34	81.19	65.76
	MCLNet [16]	ICCV-21	65.40	61.98	47.39	72.56	76.58	72.10	80.31	73.07	57.39	75.93	69.49	52.63
	SMCL [43]	ICCV-21	67.39	61.78	-	68.84	75.56	-	83.93	79.83	-	83.05	78.57	-
	FMCNet[60]	CVPR-22	66.34	62.51	-	68.15	74.09	-	89.12	84.43	-	88.38	83.86	-
	MAUM [27]	CVPR-22	71.68	68.79	-	76.97	81.94	-	87.87	85.09	-	86.95	84.34	-
	DEEN [61]	CVPR-23	74.7	71.8	-	80.3	83.3	-	91.1	85.1	-	89.5	83.4	-
	PMCM [34]	IJCAI-23	75.54	71.16	-	81.52	84.33	-	93.09	89.57	-	91.44	87.15	-
	PartMix [20]	CVPR-23	77.78	74.62	-	81.52	84.38	-	84.93	82.52	-	85.66	82.27	-
	SGIEL [20]	CVPR-23	77.12	72.33	-	82.07	82.95	-	95.35	89.98	-	97.57	91.41	-
Unsupervised	SPCL [14]	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 [13]	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
	IICS [49]	CVPR-21	14.39	15.74	8.41	15.91	24.87	22.15	9.17	9.94	6.40	9.11	9.90	6.45
	CAP [38]	AAAI-21	16.82	15.71	7.02	24.57	30.74	26.15	9.71	11.56	8.74	10.21	11.34	7.92
	Cluster Contrast [9]	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 [3]	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 [6]	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
	OTLA [37]	ECCV-22	29.9	27.1	-	29.8	38.8	-	32.9	29.7	-	32.1	28.6	-
	H2H [24]	TIP-21	30.15	29.40	-	-	-	-	23.81	18.87	-	-	-	-
	ADCA [50]	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
	DPIS [35]	ICCV-23	58.4	55.6	-	63.0	70.0	-	62.3	53.2	-	61.5	52.7	-
	CHCR [33]	TCSVT-23	59.47	59.14	-	-	-	-	69.31	64.74	-	69.96	65.87	-
	DOTLA [5]	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 [4]	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
	PGM [48]	CVPR-23	57.27	51.78	34.96	56.23	62.74	58.13	69.48	65.41	-	69.85	65.17	-
	GUR [51]	ICCV-23	63.51	61.63	47.93	71.11	76.23	72.57	73.91	70.23	58.88	75.00	69.94	56.21
	SDCL (ours)	-	**64.49**	**63.24**	**51.06**	**71.37**	**76.90**	**73.50**	**86.91**	**78.92**	**62.83**	**85.76**	**77.25**	**59.57**

Table 1. Comparison with state-of-the-arts on SYSU-MM01 and RegDB. Rank at r accuracy(%), mAP (%) and mINP (%) are reported.

					SYSU-MM01						RegDB					
	Components				All Search			Indoor Search			Visible to Infrared			Infrared to Visible		
Index	Baseline	$\mathcal{L}_{id}^{vs}+\mathcal{L}_{id}^{rs}$	CNL	CRA	$r1$	mAP	mINP	$r1$	mAP	mINP	$r1$	mAP	mINP	$r1$	mAP	mINP
1	✓				49.55	48.70	36.92	53.85	60.96	57.05	63.16	58.12	42.59	61.70	56.76	41.65
2	✓	✓			53.21	50.14	40.80	60.15	65.24	64.02	69.97	63.34	48.98	68.54	63.19	48.06
3	✓	✓	✓		55.27	54.79	41.61	64.38	71.15	67.40	76.60	67.97	51.38	75.64	68.46	50.38
4	✓	✓		✓	59.43	59.18	48.81	67.48	74.31	70.30	80.53	73.65	58.13	79.13	70.18	53.30
5	✓	✓	✓	✓	**64.49**	**63.24**	**51.06**	**71.37**	**76.90**	**73.50**	**86.91**	**78.92**	**62.83**	**85.76**	**77.25**	**59.57**

Table 2. Ablation studies on the SYSU-MM01 and RegDB. Rank at r accuracy (%), mAP (%) and mINP (%) are reported.

4.4. Ablation Study

In this subsection, we meticulously conduct ablation experiments to validate the individual efficacy of each module. The results are reported in Table 2.

Baseline denotes the augmented dual-contrastive learning framework [50] with a dual-path transformer architecture. The optimization for the network only uses deep features, *i.e.*, training the model with $\mathcal{L}_{id}^{vd}+\mathcal{L}_{id}^{rd}$.

Effectiveness of Deep and Shallow Contrastive Learning. The incorporation of $\mathcal{L}id^{vs}+\mathcal{L}id^{rs}$ yields a 4%-7% enhancement of rank-1 accuracy on SYSU-MM01 and RegDB datasets compared to the baseline. Despite shallow contrastive learning primarily optimizing the ReID model for feature learning within intra-modality, the findings underscore its capacity to augment cross-modality retrieval.

Effectiveness of CNL. Compared with the baseline, the CNL improves the performance of 6%-10% on SYSU-MM01 and RegDB datasets. The main gain is achieved by the design of capturing credible homogeneous and hetero-

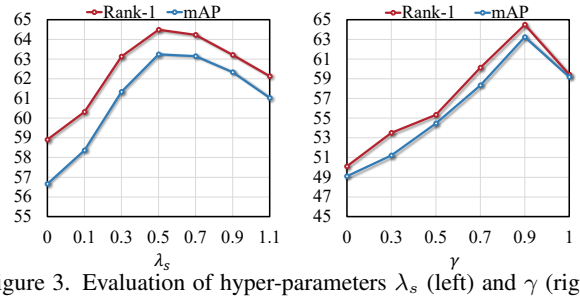


Figure 3. Evaluation of hyper-parameters λ_s (left) and γ (right). The results are based on all search mode of SYSU-MM01 dataset. Rank-1 accuracy (%) and mAP (%) are reported.

geneous neighborhoods with the shallow and deep collaborative consistency, which is a dynamic optimization, reducing the impact of label noise in clustering and enhancing the robustness against multiple discrepancies.

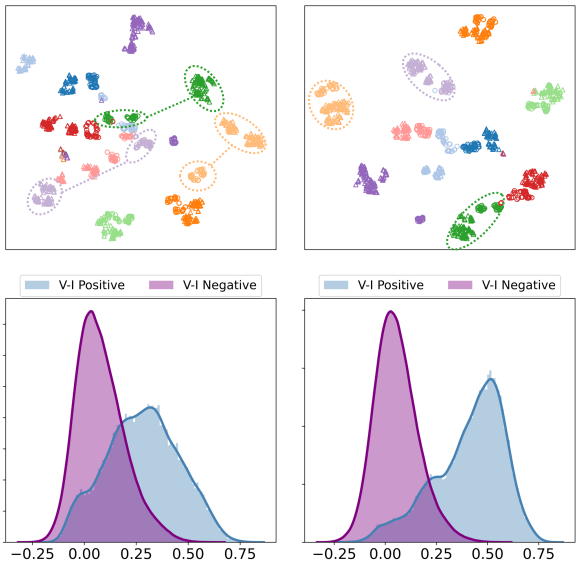
Effectiveness of CRA. In contrast to the baseline, the CRA module remarkably improves the accuracy. CRA establishes a linkage between the labels of two modalities through collaborative shallow and deep ranking relations, constituting a global association across the training set and effectively bolstering modality-invariant features.

Effectiveness of SDCL. The amalgamation of CNL and CRA notably enhances Rank-1 accuracy across diverse settings. This observation substantiates the rationale behind SDCL. CRA provides cross-modality supervision at the beginning of each training epoch in an off-line manner, while CNL dynamically learns cross-modality features during training iteration in an online manner. By combining off-line label association and online feature learning, SDCL learns better representation from different perspectives.

4.5. Further Analysis

Hyper-parameter Analysis for λ_s and γ. We explore the influence of hyper-parameters λ_s in Eq. 20 and γ in Eq. 12, Eq. 13, Eq. 16, and Eq. 17, as shown in Fig. 3. The λ_s governs the balance of $\mathcal{L}_d$ and $\mathcal{L}_s$ in deep and shallow neighbor learning. When $\lambda_s = 0$, the $\mathcal{L}_s$ is excluded. When $\lambda_{base} = 0.5$, the baseline method achieves a balance in deep and shallow neighbor learning. The γ is used to control the positive neighbor selection range in CNL. Setting $\gamma = 0$ to 0 renders neighbor selection ineffective. When $\gamma = 0.9$, the CNL significantly reinforces the cross-modality learning.

Visualization. We perform feature space (t-SNE [36] map) and similarity distribution visualization for SDCL, as presented in Fig. 4. As evident in (a) and (b), SDCL brings infrared and visible positive sample points closer together, and effectively enhances the separation of cross-modality positive/negative distributions. This attests to the efficacy of our method in addressing modality discrepancies. Additionally, the visualization of shallow and deep feature map is presented in the **supplementary materials**.



(a) Baseline (b) SDCL

Figure 4. The t-SNE (first row) and similarity distribution (second row) visualization of randomly selected identities. In the t-SNE visualization, identity is denoted by color, where circles represent the visible modality and triangles denote the infrared modality.

5. Conclusion

This paper investigates an extremely important and challenging problem, namely the unsupervised visible infrared person re-identification (US-VI-ReID) task, alleviating the reliance on expensive cross-modality annotation. To remedy the issues of multiple discrepancies, we propose a comprehensive Shallow-Deep Collaborative Learning (SDCL) framework based on the transformer architecture, incorporating Collaborative Neighbor Learning (CNL), and Collaborative Ranking Association (CRA) module. SDCL significantly facilitates the learning of robust representation, effectively countering the cross-modality discrepancy through the collaboration of shallow patch embeddings and deep modality-shared features and capturing more discriminative representations for cross-modality retrieval. Experiments on two public benchmarks substantiate that our approach surpasses existing methodologies by a substantial margin. Moreover, it even outshines certain supervised counterparts, propelling VI-ReID toward real-world deployment.

Acknowledgments

This work is partially supported by the National Natural Science Foundation of China under Grants (62071338, 62176188, 62272354, 62306215), the Key Research and Development Program of Hubei Province (2022BCA009,2022BAD175), the Special Fund of Hubei Luojia Laboratory (220100015), and the Interdisciplinary Innovative Talents Foundation from Renmin Hospital of Wuhan University.

References

[1] Cuiqun Chen, Mang Ye, Meibin Qi, Jingjing Wu, Yimin Liu, and Jianguo Jiang. Saliency and granularity: Discovering temporal coherence for video-based person re-identification. *IEEE TCSVT*, 2022. 1

[2] Cuiqun Chen, Mang Ye, Meibin Qi, and Bo Du. Sketchtrans: Disentangled prototype learning with transformer for sketch-photo recognition. *IEEE TPAMI*, 2023. 1

[3] Hao Chen, Benoit Lagadec, and Francois Bremond. Ice: Inter-instance contrastive encoding for unsupervised person re-identification. In *ICCV*, 2021. 2, 7

[4] 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 *ACM MM*, 2023. 2, 7

[5] 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. *ACM MM*, 2023. 2, 7

[6] Yoonki Cho, Woo Jae Kim, Seunghoon Hong, and Sung-Eui Yoon. Part-based pseudo label refinement for unsupervised person re-identification. In *CVPR*, 2022. 7

[7] Zhenyu Cui, Jiahuan Zhou, and Yuxin Peng. Dma: Dual modality-aware alignment for visible-infrared person re-identification. *IEEE TIFS*, 2024. 1

[8] Yongxing Dai, Jun Liu, Yan Bai, Zekun Tong, and Ling-Yu Duan. Dual-refinement: Joint label and feature refinement for unsupervised domain adaptive person re-identification. *IEEE TIP*, 2021. 2

[9] Zuozhuo Dai, Guangyuan Wang, Siyu Zhu, Weihao Yuan, and Ping Tan. Cluster contrast for unsupervised person re-identification. arxiv 2021. *arXiv preprint arXiv:2103.11568*, 2021. 2, 7

[10] 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 *KDD*, 1996. 6

[11] Jiawei Feng, Ancong Wu, and Wei-Shi Zheng. Shape-erased feature learning for visible-infrared person re-identification. In *CVPR*, 2023. 2

[12] Yajun Gao, Tengfei Liang, Yi Jin, Xiaoyan Gu, Wu Liu, Yidong Li, and Congyan Lang. Mso: Multi-feature space joint optimization network for rgb-infrared person re-identification. In *ACM MM*, 2021. 1, 6, 7

[13] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. *arXiv preprint arXiv:2001.01526*, 2020. 7

[14] Yixiao Ge, Feng Zhu, Dapeng Chen, Rui Zhao, et al. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. *NeurIPS*, 2020. 7

[15] Jianyang Gu, Weihua Chen, Hao Luo, Fan Wang, Hao Li, Wei Jiang, and Weijie Mao. Multi-view evolutionary training for unsupervised domain adaptive person re-identification. *IEEE TIFS*, 2022. 1

[16] Xin Hao, Sanyuan Zhao, Mang Ye, and Jianbing Shen. Cross-modality person re-identification via modality confusion and center aggregation. In *ICCV*, 2021. 6, 7

[17] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *ICCV*, 2021. 6

[18] Wenke Huang, Mang Ye, Zekun Shi, and Bo Du. Generalizable heterogeneous federated cross-correlation and instance similarity learning. *IEEE TPAMI*, 2023. 1

[19] Wenke Huang, Mang Ye, Zekun Shi, Guancheng Wan, He Li, Bo Du, and Qiang Yang. A federated learning for generalization, robustness, fairness: A survey and benchmark. *arXiv*, 2023. 1

[20] 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 *CVPR*, 2023. 1, 7

[21] Diangang Li, Xing Wei, Xiaopeng Hong, and Yihong Gong. Infrared-visible cross-modal person re-identification with an x modality. In *AAAI*, 2020. 1

[22] Shuang Li, Jiaxu Leng, Ji Gan, Mengjingcheng Mo, and Xinbo Gao. Shape-centered representation learning for visible-infrared person re-identification. *arXiv preprint arXiv:2310.17952*, 2023. 1

[23] Shuang Li, Fan Li, Jinxing Li, Huafeng Li, Bob Zhang, Dapeng Tao, and Xinbo Gao. Logical relation inference and multiview information interaction for domain adaptation person re-identification. *IEEE TNNLS*, 2023. 1

[24] Wenqi Liang, Guangcong Wang, Jianhuang Lai, and Xiaohua Xie. Homogeneous-to-heterogeneous: Unsupervised learning for rgb-infrared person re-identification. *IEEE TIP*, 2021. 2, 3, 7

[25] Shan Lin, Chang-Tsun Li, and Alex C. Kot. Multi-domain adversarial feature generalization for person re-identification. *IEEE TIP*, 2021. 1

[26] Haojie Liu, Daoxun Xia, and Wei Jiang. Towards homogeneous modality learning and multi-granularity information exploration for visible-infrared person re-identification. *IEEE Journal of Selected Topics in Signal Processing*. 7

[27] 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 *CVPR*, 2022. 1, 7

[28] Yan Lu, Yue Wu, Bin Liu, Tianzhu Zhang, Baopu Li, Qi Chu, and Nenghai Yu. Cross-modality person re-identification with shared-specific feature transfer. In *CVPR*, 2020. 1

[29] Zefeng Lu, Ronghao Lin, and Haifeng Hu. Modality and camera factors bi-disentanglement for nir-vis object re-identification. *IEEE TIFS*, 2023.

[30] Chuanchen Luo, Chunfeng Song, and Zhaoxiang Zhang. Learning to adapt across dual discrepancy for cross-domain person re-identification. *IEEE TPAMI*, 2022. 1, 5

[31] Hao Luo, Pichao Wang, Yi Xu, Feng Ding, Yanxin Zhou, Fan Wang, Hao Li, and Rong Jin. Self-supervised pre-training for transformer-based person re-identification. *arXiv preprint arXiv:2111.12084*, 2021. 6

[32] Dat Tien Nguyen, Hyung Gil Hong, Ki Wan Kim, and Kang Ryoung Park. Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors*, 2017. 6

[33] Zhiqi Pang, Chunyu Wang, Lingling Zhao, Yang Liu, and Gaurav Sharma. Cross-modality hierarchical clustering and refinement for unsupervised visible-infrared person re-identification. *IEEE TCSVT*, 2023. 2, 3, 7

[34] Zhihao Qian, Yutian Lin, and Bo Du. Visible-infrared person re-identification via patch-mixed cross-modality learning. *IJCAI*, 2023. 1, 7

[35] Jiangming Shi, Yachao Zhang, Xiangbo Yin, Yuan Xie, Zhizhong Zhang, Jianping Fan, Zhongchao Shi, and Yanyun Qu. Dual pseudo-labels interactive self-training for semi-supervised visible-infrared person re-identification. In *ICCV*, 2023. 6, 7

[36] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 2008. 8

[37] 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. *ECCV*, 2022. 2, 3, 6, 7

[38] Menglin Wang, Baisheng Lai, Jianqiang Huang, Xiaojin Gong, and Xian-Sheng Hua. Camera-aware proxies for unsupervised person re-identification. In *AAAI*, 2021. 7

[39] Yiming Wang, Guanqiu Qi, Shuang Li, Yi Chai, and Huafeng Li. Body part-level domain alignment for domain-adaptive person re-identification with transformer framework. *IEEE TIFS*, 2022. 1

[40] Yadi Wang, Hongyun Zhang, Duoqian Miao, and Witold Pedrycz. Multi-granularity re-ranking for visible-infrared person re-identification. *CAAI Transactions on Intelligence Technology*, 2023. 1

[41] Zhixiang Wang, Zheng Wang, Yinqiang Zheng, Yung-Yu Chuang, and Shin'ichi Satoh. Learning to reduce dual-level discrepancy for infrared-visible person re-identification. In *CVPR*, 2019. 1

[42] 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

[43] Ziyu Wei, Xi Yang, Nannan Wang, and Xinbo Gao. Syncretic modality collaborative learning for visible infrared person re-identification. In *ICCV*, 2021. 7

[44] Ancong Wu, Wei-Shi Zheng, and Jian-Huang Lai. Robust depth-based person re-identification. *IEEE TIP*, 2017. 1

[45] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. Rgb-infrared cross-modality person re-identification. In *ICCV*, 2017. 1, 6

[46] Dongming Wu, Mang Ye, Gaojie Lin, Xin Gao, and Jianbing Shen. Person re-identification by context-aware part attention and multi-head collaborative learning. *IEEE TIFS*, 2022. 1

[47] Qiong Wu, Pingyang Dai, Jie Chen, Chia-Wen Lin, Yongjian Wu, Feiyue Huang, Bineng Zhong, and Rongrong Ji. Discover cross-modality nuances for visible-infrared person re-identification. In *CVPR*, 2021. 1, 7

[48] Zesen Wu and Mang Ye. Unsupervised visible-infrared person re-identification via progressive graph matching and alternate learning. In *CVPR*, 2023. 2, 3, 6, 7

[49] Shiyu Xuan and Shiliang Zhang. Intra-inter camera similarity for unsupervised person re-identification. In *CVPR*, 2021. 2, 7

[50] Bin Yang, Mang Ye, Jun Chen, and Zesen Wu. Augmented dual-contrastive aggregation learning for unsupervised visible-infrared person re-identification. In *ACM MM*, 2022. 1, 2, 3, 6, 7

[51] Bin Yang, Jun Chen, and Mang Ye. Towards grand unified representation learning for unsupervised visible-infrared person re-identification. In *ICCV*, 2023. 2, 6, 7

[52] Bin Yang, Jun Chen, and Mang Ye. Top-k visual tokens transformer: Selecting tokens for visible-infrared person re-identification. In *ICASSP*, 2023. 1

[53] Mang Ye, Xiangyuan Lan, Zheng Wang, and Pong C Yuen. Bi-directional center-constrained top-ranking for visible thermal person re-identification. *IEEE TIFS*, pages 407–419, 2019. 1, 6

[54] Mang Ye, Jianbing Shen, David J Crandall, Ling Shao, and Jiebo Luo. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In *ECCV*, 2020. 1, 6, 7

[55] Mang Ye, Cuiqun Chen, Jianbing Shen, and Ling Shao. Dynamic tri-level relation mining with attentive graph for visible infrared re-identification. *IEEE TIFS*, 2021. 1, 2

[56] Mang Ye, Weijian Ruan, Bo Du, and Mike Zheng Shou. Channel augmented joint learning for visible-infrared recognition. In *ICCV*, 2021. 2, 6, 7

[57] 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 TPAMI*, 2021. 1, 6, 7

[58] Mang Ye, Zesen Wu, Cuiqun Chen, and Bo Du. Channel augmentation for visible-infrared re-identification. *IEEE TPAMI*, 2023. 1

[59] Mingyang Zhang, Yang Xiao, Fu Xiong, Shuai Li, Zhiguo Cao, Zhiwen Fang, and Joey Tianyi Zhou. Person re-identification with hierarchical discriminative spatial aggregation. *IEEE TIFS*, 2022. 1

[60] Qiang Zhang, Changzhou Lai, Jianan Liu, Nianchang Huang, and Jungong Han. Fmcnet: Feature-level modality compensation for visible-infrared person re-identification. In *CVPR*, 2022. 7

[61] Yukang Zhang and Hanzi Wang. Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification. In *CVPR*, 2023. 1, 7

[62] Yiyuan Zhang, Yuhao Kang, Sanyuan Zhao, and Jianbing Shen. Dual-semantic consistency learning for visible-infrared person re-identification. *IEEE TIFS*, 2022. 1

[63] Yafei Zhang, Yongzeng Wang, Huafeng Li, and Shuang Li. Cross-compatible embedding and semantic consistent feature construction for sketch re-identification. In *ACM MM*, 2022. 1

[64] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015. 3