Towards Grand Unified Representation 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/GUR

Abstract

Unsupervised learning visible-infrared person re-identification (USL-VI-ReID) is an extremely important and challenging task, which can alleviate the issue of expensive cross-modality annotations. Existing works focus on handling the cross-modality discrepancy under unsupervised conditions. However, they ignore the fact that USL-VI-ReID is a cross-modality retrieval task with the hierarchical discrepancy, i.e., camera variation and modality discrepancy, resulting in clustering inconsistencies and ambiguous cross-modality label association. To address these issues, we propose a hierarchical framework to learn grand unified representation (GUR) for USL-VI-ReID. The grand unified representation lies in two aspects: 1) GUR adopts a bottom-up domain learning strategy with a cross-memory association embedding module to explore the information of hierarchical domains, i.e., intra-camera, inter-camera, and inter-modality domains, learning a unified and robust representation against hierarchical discrepancy. 2) To unify the identities of the two modalities, we develop a cross-modality label unification module that constructs a cross-modality affinity matrix as a bridge for propagating labels between two modalities. Then, we utilize the homogeneous structure matrix to smooth the propagated labels, ensuring that the label structure within one modality remains unchanged. Extensive experiments demonstrate that our GUR framework significantly outperforms existing USL-VI-ReID methods, and even surpasses some supervised counterparts.

1. Introduction

Person re-identification (ReID) aims at matching the same person images captured by non-overlapping cameras [13, 16]. This technology has been widely investigated due to its significance for social security. Most existing ReID models concentrate on the single-modality image matching task with RGB images captured by visible cameras. However, visible cameras cannot capture enough information under poor illumination conditions [54]. Hence, visible infrared person re-identification (VI-ReID) has emerged to match person images captured by visible and infrared cameras for the 24-hour surveillance system [35, 42, 55].

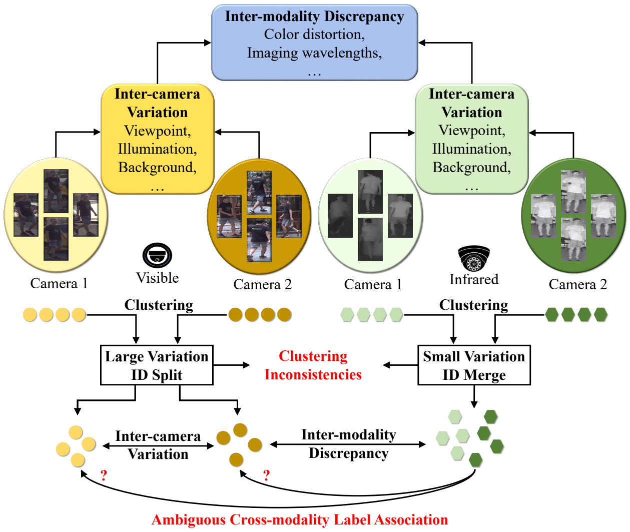

Figure 1. Illustration of hierarchical discrepancy in USL-VI-ReID with two cameras within each modality as an example. Circles and hexagons represent the sample points of the same person from infrared and visible modalities, respectively. Different colors represent different cameras and modalities. The inter-camera variation and inter-modality discrepancy collectively result in clustering inconsistencies and ambiguous cross-modality label association.

Existing VI-ReID methods have achieved remarkable performance with deep learning methods [56, 51, 49, 50]. However, the success mainly profits from supervised learning over massive human-labeled data, which is more time-consuming and expensive than manual annotations in single-modality ReID [21, 43]. Recently, unsupervised learning visible infrared person re-identification (USL-VI-

[†] Corresponding Author.

ReID) [21, 43, 31] has been proposed to alleviate the issue of expensive cross-modality annotations.

In USL-VI-ReID, unsupervised settings and hierarchical discrepancies in both inter-camera and inter-modality make it more challenging and different from unsupervised single-modality ReID. The inter-camera variation and inter-modality discrepancy collectively form the hierarchical discrepancy, which complicates the learning of the USL-VI-ReID model, *e.g*., leading to clustering inconsistencies and ambiguous cross-modality label association, as illustrated in Fig. 1. The variations between the cameras of the two modalities are different. Visible and infrared cameras have different sensitivities to light. In general, RGB cameras are more susceptible to light and other factors compared with IR cameras. Large variations may make identities split and small variations may enable identities to merge, leading to inconsistent cluster numbers of the two modalities and significantly increasing the difficulty of cross-modality label association. More importantly, the hierarchical discrepancy is not simply camera variation plus modality discrepancy, but a complex misalignment of features and cross-modality labels, hindering the retrieval of the same person across different modalities. We will show that our approach significantly alleviates clustering inconsistencies in the experiments. For better cross-modality retrieval performance, it is desirable to handle the aforementioned hierarchical discrepancy. Existing methods [21, 43, 31] for USL-VI-ReID usually focus on solving the problem of modality discrepancy. However, they ignored the hierarchical discrepancy, hindering further improvement.

To handle the hierarchical discrepancy in USL-VI-ReID, we put forward a novel grand unified representation (GUR) learning framework to explore the information of hierarchical domains. GUR adopts a bottom-up domain learning strategy with a cross-memory association embedding (CAE) and cross-modality label unification (CLU) module. The bottom-up domain learning strategy consists of *intra-camera training*, *inter-camera* and *inter-modality training*. At the inter-camera and inter-modality training stage, a CAE module is developed to calculate the association probability embedding between a pedestrian image and each memory item of one domain, and collect the association probabilities of camera or modality of all domains as the unified probability embedding for clustering. To further associate the cross-modality identities, we introduce a CLU module to construct a top-k heterogeneous affinity matrix as the bridge for propagating labels between two modalities and use the homogeneous structure matrix to smooth the propagated labels, ensuring that the label structure within one modality remains unchanged. Finally, with the above bottom-up domain learning strategy with the CAE module and CLU module, our method learns a unified representation, achieving both camera- and modality-invariant properties.

The main contributions are summarized as follows:

- We propose a novel unsupervised learning framework that adopts a bottom-up domain learning strategy with cross-memory association embedding. This enables the model to learn unified representation which is robust against hierarchical discrepancy.

- We design a cross-modality label unification module to propagate and smooth labels between two modalities with heterogeneous affinity matrix and homogeneous structure matrix, respectively, unifying the identities across the two modalities.

- Extensive experiments on the SYSU-MM01 and RegDB datasets demonstrate that our GUR framework significantly outperforms existing USL-VI-ReID methods, and even surpasses some supervised counterparts, further narrowing the gap between supervised and unsupervised VI-ReID.

2. Related Work

2.1. Supervised Visible-Infrared Person ReID

VI-ReID has received extensive attention due to its ability to search out the same person under poor illumination conditions at night. Many works [46, 38, 50, 52, 45, 1, 48, 44, 47, 51, 17, 54, 37, 39] have been developed to overcome the modality discrepancy between infrared and visible cameras. Ye *et al*. [49] proposed Channel exchangeable Augmentation (CA) to homogeneously generate color-irrelevant images by randomly exchanging the color channels, improving the robustness against color variations. Liu *et al*. [22] proposed the Memory-augmented Unidirectional Metric (MAUM) learning method to enforce explicit cross-modality association with two unidirectional metrics. To compensate for the missing modality-specific information in the feature level, Zhang *et al*. [54] directly generated those missing modality-specific features of one modality from existing modality-shared features of the other modality.

These methods have achieved surprising performance with supervised learning over massive human-labeled data, which is more time-consuming and expensive than the manual annotations in single-modality ReID. Differently, our proposed framework trains a VI-ReID model without any identity annotations, alleviating the issue of expensive cross-modality annotations.

2.2. Unsupervised Single-Modality Person ReID

Existing unsupervised single-modality person ReID methods can be divided into pseudo-label-based methods

[14, 11, 25, 2, 40, 41, 15, 32] and translation-based methods [58, 36, 4, 3, 60], where the former achieve better performance. Dai *et al.* [7] proposed a cluster contrast that computes contrast loss at the cluster level to solve the problem of inconsistency in the updating progress of each cluster. A camera-aware proxy assisted learning method was introduced in [32] to deal with the large intra-ID variance caused by the change of camera views. To address the challenge of distribution discrepancy among cameras, Xuan *et al.* [40, 41] decomposed the sample similarity computation into intra-camera and inter-camera computations.

Although the above methods have a promising performance on single-modality unsupervised ReID, the large cross-modality discrepancy prevents them from solving the USL-VI-ReID problem.

2.3. Unsupervised Visible-Infrared Person ReID

The existing methods [43, 21, 31] for USL-VI-ReID focus on reducing modality gap. ADCA [43] proposed an Augmented Dual-Contrastive Aggregation (ADCA) learning framework to learn the inter-modality person representation and associate positive cross-modality identities under purely unsupervised conditions. H2H [21] designed a homogeneous learning and heterogeneous learning method to solve the USL-VI-ReID task using the Market-1501 dataset [57] as an extra labeled RGB dataset for pre-training. OTLA [31] developed an optimal-transport strategy trying to assign pseudo labels from visible to infrared modality.

These methods were initial attempts at the USL-VI-ReID task. However, they ignored the hierarchical discrepancy, limiting the discriminability of features against camera variations and modality discrepancy. In contrast to the prior works, we simultaneously consider both aspects of hierarchical discrepancy, *i.e.*,inter-camera variation and inter-modality discrepancy, significantly improving the cross-modality retrieval performance.

3. Proposed Method

3.1. Overview

We propose a grand unified representation (GUR) learning framework to address the problem of hierarchical discrepancy, as shown in Fig. 2. The GUR framework contains a bottom-up domain learning strategy with a cross-memory association embedding module and a cross-modality label unification module.

Bottom-up domain learning strategy has three training stages, *i.e.*, *intra-camera training*, *inter-camera* and *inter-modality training*. In each stage, the augmented dual-contrastive (ADC) learning [43] is conducted for pseudo-label-based unsupervised learning. We extract features for all training samples and then use DBSCAN clustering algorithm [10] to assign pseudo-labels. In the intra-camera training stage, the ADC is executed alternately in each camera domain separately via clustering the intra-camera similarity. In the inter-camera training phase, the **cross-memory association embedding (CAE)** module at the camera level calculates the association embedding of persons with each camera memory for inter-camera (intra-modality) DBSCAN clustering (*i.e.*, clustering the data of each modality separately). Then, ADC with two modality-specific memories is performed to learn camera-invariant features within each modality. Similarly, during the inter-modality training, the CAE module at the modality level computes the association embedding of pedestrians with each modality memory for inter-modality clustering (*i.e.*, simultaneously input all data into DBSCAN for clustering without considering domains), and the ADC with a modality-shared memory is conducted to learn the final unified features. From intra-camera to inter-modality training, the model progressively captures camera-invariant and modality-invariant features. Three stages are executed in an alternate manner during one training epoch. The detailed figure is shown in **supplementary materials**. To further ensure the semantic consistency of the two modality labels, we insert the **cross-modality label unification (CLU)** module between intra-modality and inter-modality. With the above modules, GUR learns unified representations for cross-modality retrieval, which are robust to the hierarchical discrepancy, *i.e.*, inter-camera variation and inter-modality discrepancy.

3.2. Preliminary

To facilitate the description of our method, we first revisit ADC [43] learning. For convenience, we omit the equations of the augmented stream.

Memory Initialization. At the beginning of each training iteration, we store each cluster's representation in infrared and visible memory $M^i = \left[m_1^i, \ldots, m_K^i\right], M^v = [m_1^v, \ldots, m_L^v]$, respectively, by the following equations:

$$m_k^i = \frac{1}{\left|\mathcal{H}_k^i\right|} \sum_{u_n^i \in \mathcal{H}_k^i} u_n^i, \tag{1}$$

$$m_l^v = \frac{1}{\left|\mathcal{H}_l^v\right|} \sum_{u_m^v \in \mathcal{H}_l^v} u_m^v, \tag{2}$$

where u_n^i and u_m^v denote the corresponding features extracted by the infrared and visible feature extractor f_θ^i and f_θ^v. $\mathcal{H}_{k(l)}^{i(v)}$ is the k or l-th cluster set in infrared or visible modality according to the clustering results of DBSCAN[10]. $|\cdot|$ represents the number of instances per cluster. During training, we update the two modality-specific memories by a momentum updating strategy [7].

Loss Function. We update the feature extractor by ClusterNCE [7] loss within infrared and visible modality, which

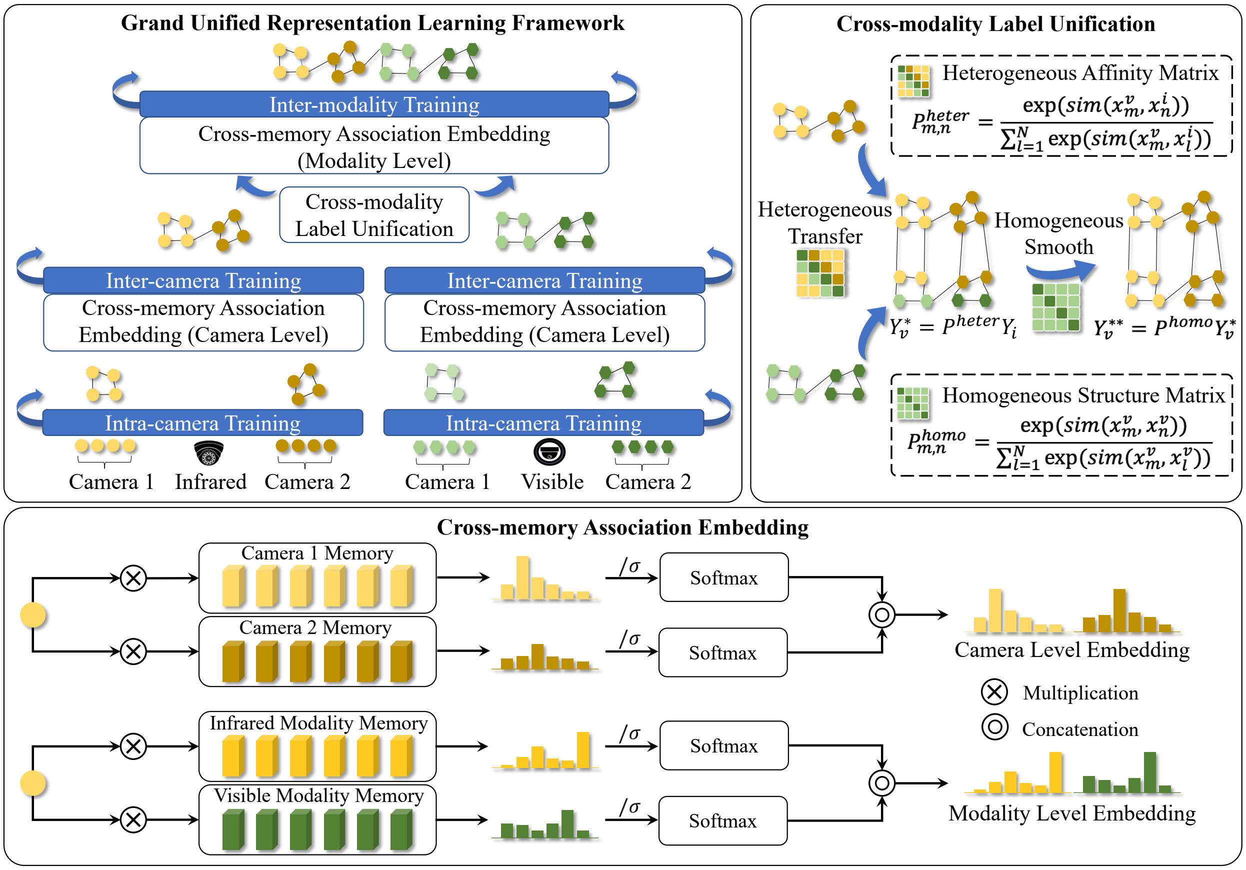
Grand Unified Representation Learning Framework
Inter-modality Training
Cross-memory Association Embedding (Modality Level)
Cross-modality Label Unification
Inter-camera Training
Cross-memory Association Embedding (Camera Level)
Inter-camera Training
Cross-memory Association Embedding (Camera Level)
Intra-camera Training
Intra-camera Training
Camera 1
Infrared
Camera 2
Camera 1
Visible
Camera 2
Cross-modality Label Unification
Heterogeneous Affinity Matrix
$P^{heter}_{m,n} = \frac{\exp(sim(x^v_m, x^i_n))}{\sum_{l=1}^{N} \exp(sim(x^v_m, x^i_l))}$
Heterogeneous Transfer
Homogeneous Smooth
$Y^*_v = P^{heter} Y_i$
$Y^{**}_v = P^{homo} Y^*_v$
Homogeneous Structure Matrix
$P^{homo}_{m,n} = \frac{\exp(sim(x^v_m, x^v_n))}{\sum_{l=1}^{N} \exp(sim(x^v_m, x^v_l))}$
Cross-memory Association Embedding
Camera 1 Memory
Camera 2 Memory
Infrared Modality Memory
Visible Modality Memory
$/\sigma$
$/\sigma$
$/\sigma$
$/\sigma$
Softmax
Softmax
Softmax
Softmax
Camera Level Embedding
Multiplication
Concatenation
Modality Level Embedding

Figure 2. Illustration of grand unified representation learning framework with two cameras within each modality as an example. Circles and hexagons represent the sample points of the same person from infrared and visible modalities, respectively. Different colors represent different cameras and modalities. It comprises a bottom-up domain learning strategy with a cross-memory association embedding module and a cross-modality label unification module, which contains intra-camera training, inter-camera training, and inter-modality training. Cross-memory association embedding module calculates the association embedding using all camera or modality memories to generate reliable pseudo labels for ADC learning. Cross-modality label unification module unifies the pseudo labels of the infrared and visible modality, ensuring the semantic consistency of the two modality labels.

can be calculated as:

$$L_{q_i} = -\log \frac{\exp\left(q_i \cdot m^i_+ / \tau\right)}{\sum_{k=0}^{K} \exp\left(q_i \cdot m^i_k / \tau\right)}, \tag{3}$$

$$L_{q_v} = -\log \frac{\exp\left(q_v \cdot m^v_+ / \tau\right)}{\sum_{l=0}^{L} \exp\left(q_v \cdot m^v_l / \tau\right)}, \tag{4}$$

where m_+ is the positive cluster representation and the τ is a temperature hyper-parameter. q_i and q_v are query instance features extracted by f^i_θ and f^v_θ, respectively.

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

$$L_{overall} = L_{q_i} + L_{q_v}. \tag{5}$$

3.3. Cross-memory Association Embedding

The CAE module calculates the association probability between a pedestrian feature and each memory item within one domain and collects the association probabilities of all cameras or modalities as the unified probability embedding for clustering. The rationale is that samples belonging to the same identity should have a similar distribution of association probability produced by each memory [9, 27]. This distribution conceptually represents the affinity of the image with each cluster features in each domain, and the final embedding consisting of the distribution of all domains is robust against hierarchical domain variations. [9, 27].

Given a memory M_n as a probability mapping matrix, the process of calculating the association embedding $e\left(q|M_n\right)$ of instance feature q can be represented as:

$$p(y|q, M_n) = \frac{\exp(q \cdot m_y / \sigma)}{\sum_{c=1}^{C} \exp\left(q \cdot m_c / \sigma\right)}, \tag{6}$$

$$e\left(q|M_n\right) = [p(1|q, M_n), p(2|q, M_n), \cdots, p(C|q, M_n)], \tag{7}$$

where m_y is the memory feature of label y. $p(c|q, M_n)$ is the association probability of the instance feature q at class c. σ is a temperature hyper-parameter.

We concatenate all association embeddings from N memories as the cross-memory association embedding, *i.e*.,

$$E\left(q\right) = \left[e\left(q|M_1\right), e\left(q|M_2\right), \cdots, e\left(q|M_N\right)\right], \tag{8}$$

where $E\left(q\right)$ is the embedding for clustering with DBSCAN[10]. In the inter-camera training, the probability mapping matrix M_n is the cluster memory in the camera n domain produced by intra-camera training. Similarly, in the stage of inter-modality training, the probability mapping matrix M_n is the cluster memory provided by intra-modality (inter-camera) training. Note that in visual surveillance, the camera label n is naturally available for each image following popular setting [53, 32, 2, 41, 13], since it is straightforward to know by which camera an image is captured in a camera network. Cross-memory association embeddings at the camera level and modality level enable the GUR to form a hierarchical learning framework that progressively learns camera-invariant and modality-invariant representations.

What is σ doing? The main consideration of adding σ is that it can control the attention to hard negative samples. Small σ penalizes much more on the hardest negative samples, making a large difference between the probability of positive and negative sample pairs, and the embedding space is likely to be more uniform [29, 33]. When σ approaches 0, the probability embedding is likely to be a one-hot code and has less tolerance to potential positive samples [29, 33]. Large σ makes the probability smooth and less sensitive to the hard negative samples, and the hardness-aware property disappears as the σ approaches $+\infty$. The cross-memory association embedding meets a uniformity-tolerance dilemma. To get a better representation for clustering, we set σ to different values to evaluate the effect in the experiments.

Discussion. In contrast to the inter-camera training in [40, 41], which computes the embedding for clustering by concatenating the classification scores from different classifiers, we calculate the embedding with the memory in different domains (camera or modality) to seek reliable clustering across domains. The major advantages are two-fold: 1) Our association embedding is produced by the memory in Cluster Contrast [7], which does not require additional classifiers and is a tight coupling with clustering algorithms and contrastive learning. It is a mutual reinforcement. 2) We introduce a temperature hyper-parameter σ in Eq 6 to control the concern on hard negative samples. By means of σ, the CAE module can yield more discriminative and robust representations for clustering.

3.4. Cross-modality Label Unification

Through the bottom-up domain learning strategy with the CAE module, we can capture the more robust embedding for clustering across different cameras and modalities. However, the embedding still has a strong implicit correlation with the modality, which negatively impacts the generation of cross-modality pseudo labels. In response, we develop a cross-modality label unification (CLU) module.

The CLU module is based on two rationales: 1) Similar features across modalities should have the same identity. 2) Features in the same cluster within one modality should share the same identity. Accordingly, the CLU module contains two processes, *i.e*., heterogeneous transfer and homogeneous structure smooth. In heterogeneous transfer, a top-k heterogeneous affinity matrix is constructed as the bridge for propagating labels between two modalities. Then, the homogeneous structure matrix is used to smooth the propagated labels, ensuring that the label structure of the modality remains unchanged. In our work, we propagate the pseudo labels of the infrared modality to the visible modality.

Let $X_i = \{x_1^i, x_2^i, \cdots, x_N^i\}$ represent the infrared images with N instances. $X_v = \{x_1^v, x_2^v, \cdots, x_M^v\}$ denote the visible sets with M instances, respectively.

Given an instance pair $< x_m, x_n >$, we compute the similarity by:

$$sim(x_m, x_n) = \frac{f_\theta(x_m) \cdot f_\theta(x_n)}{||f_\theta(x_m)||_2 ||f_\theta(x_n)||_2}, \tag{9}$$

where f_θ is the feature extractor. The heterogeneous affinity matrix $P^{heter} \in \mathbb{R}^{M \times N}$ is formed by:

$$P_{m,n}^{heter} = \frac{\exp(sim(x_m^v, x_n^i))}{\sum_{l=1}^{N} \exp(sim(x_m^v, x_l^i))}. \tag{10}$$

We only keep the k-max values in each row of P to construct a top-k affinity matrix. Then, we transfer the infrared pseudo labels to the visible instances, which can be written as:

$$Y_v^* = P^{heter} Y_i, \tag{11}$$

where Y_i is the infrared pseudo label matrix with $Y_{m,n} = 1$ if x_m^i is labeled as $y_m = n$, otherwise $Y_{m,n} = 0$. It utilizes the heterogeneous affinity matrix of visible and infrared to weight the infrared pseudo labels and determine which category the visible labels should belong to, which can be seen as a weighted voting process. Then, we convert the weight label matrix Y_v^* to the form of one-hot code by setting the column with the largest value in each row to 1 and the rest to 0.

To further increase the credibility of the cross-modality label, we leverage a homogeneous structure matrix to smooth propagated labels, as homogeneous similarities without the interference of cross-modality variations are more reliable than heterogeneous similarities. The homogeneous structure matrix $P^{homo} \in \mathbb{R}^{M \times M}$ is defined by:

$$P_{mn}^{homo} = \frac{\exp(sim(x_m^v, x_n^v))}{\sum_{l=1}^{N} \exp(sim(x_m^v, x_l^v))}. \tag{12}$$

Similarly, we keep the k-max values in each row of P^{homo}. The process of homogeneous structure smooth is formulated as follows:

$$Y_v^{**} = P^{homo} Y_v^{*}, \tag{13}$$

where Y_v^{**} is the final unified label matrix. In Y_v^{**}, the column number of the maximum value in each row is the class label of samples.

Discussion. Heterogeneous transfer and homogeneous structure smooth in CLU are essentially two weighted voting processes, which is distinguished from OTLA [31] and ADCA [43]. OTLA [31] smooths the pseudo label by an assumption of approximately same number of infrared images within each generated pseudo label, which has a limited application under unbalanced conditions. ADCA [43] associates positive cross-modality identities with a count priority selection strategy but misses the smoothing technique to reduce cross-modality label noise. While our CLU module propagates and smooths cross-modality labels with heterogeneous transfer and homogeneous structure smooth, which is more flexible and reliable without any strong assumption.

4. Experiments

4.1. Datasets and Evaluation Protocol

We evaluate our proposed GUR on two widely-used visible infrared person ReID datasets, *i.e.*, SYSU-MM01 [38] and RegDB [24].

SYSU-MM01 is a large-scale visible VI-ReID dataset consisting of 2 infrared and 4 visible cameras. Specially, SYSU-MM01 contains 395 identities including 22258 visible images and 11909 near-infrared images for training. In testing, the query set contains 96 persons with 3803 infrared images, and the gallery set has 301 randomly selected visible images. Meanwhile, we adopt *all-search* and *indoor-search* modes [51] for evaluation.

RegDB is collected by one visible and one infrared camera in a dual-camera system. RegDB has 412 persons, and each person contains 10 infrared and 10 visible images. We randomly select 206 persons for training and another 206 identities for testing with two modes, *i.e.*, thermal to visible and visible to thermal.

Evaluation Protocol. We adopt the cumulative matching characteristics (CMC), mean average precision (mAP) and mean inverse negative penalty (mINP) [51] as the evaluation metrics. Following existing methods [48, 50, 51], we perform ten trials of the gallery set selection, and calculate the average performance to obtain stable performance.

4.2. Implementation Details

The proposed framework is implemented in PyTorch. GUR adopts the feature extractor in ADC [43] as the backbone network, which consists of shallow modality-specific layers and shared layers (ResNet50 [19]). We initialize the feature extractor with ImageNet-pretrained weights [8]. The features of the global average pooling layer are used to calculate the cosine similarity for retrieval. At the start of each stage, DBSCAN [10] is performed to generate pseudo labels. During training, person images are resized to 288×144. 16 identities and 16 instances for each identity are sampled for one batch. We adopt horizontal flipping, random crop, and random erasing for data argumentation. In addition, we utilize Channel Augmentation (CA) [49] in the augmented visible stream. Adam optimizer is adopted to train the model with the initial learning rate of $3.5e-4$. The learning rate is reduced to 1/10 of its previous value every 20 epochs. The model is trained in total of 50 epochs. The CLU module is added in the last 20 epochs. The σ in Eq 6 is set to 0.05. The other settings of dual-contrastive learning follow [43].

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

We report 19 supervised and 12 unsupervised methods for comparison. Some advanced unsupervised methods,*i.e.* IICS[40], CAP [32], and ICE [2], also use the camera label for training. Since the RegDB dataset has only one visible and one infrared camera, there are only intra-modality and inter-modality training on the RegDB task. We also report the results of GUR without using camera information on SYSU-MM01 for comparison, in which we remove the intra-camera training and the CAE at the camera level and directly perform the intra-modality (inter-camera training in Figure 2), inter-modality training with the CAE at the modality level, and CLU module.

Comparison with Unsupervised Methods. As reported in Table 1, the performance of our method surpasses current leading unsupervised methods. More precisely, our GUR achieves 63.51% and 73.91% rank-1 accuracy on SYSU-MM01 (all search) and RegDB (visible to infrared), respectively. It significantly outperforms ADCA [43] and H2H [21] by about 20% and 30% rank-1 accuracy on SYSU-MM01 and RegDB datasets. Note that our GUR also achieves the best accuracy without the camera labels compared with previous unsupervised methods. ADCA, H2H, and OTLA focus on solving the problem of modality discrepancy. However, the neglect of hierarchical discrepancy limits further improvement. Our method employs a more reasonable bottom-up domain learning framework and CLU module, ensuring robustness against the hierarchical dis-

			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	Zero-Padding [38]	ICCV-17	14.80	15.95	-	20.58	26.92	-	17.75	18.90	-	16.63	17.82	-
	eBDTR [48]	TIFS-19	27.82	28.42	-	32.46	42.46	-	34.62	33.46	-	34.21	32.49	-
	HSME [18]	AAAI-19	20.68	23.12	-	-	-	-	50.85	47.00	-	50.15	46.16	-
	D^{2}RL [34]	CVPR-19	28.9	29.2	-	-	-	-	43.4	44.1	-	-	-	-
	AlignGAN [30]	ICCV-19	42.4	40.7	-	45.9	54.3	-	57.9	53.6	-	56.3	53.4	-
	X-Modal [20]	AAAI-20	49.9	50.7	-	-	-	-	62.21	60.18	-	-	-	-
	Hi-CMD [6]	CVPR-20	34.9	35.9	-	-	-	-	70.93	66.04	-	-	-	-
	cm-SSFT* [23]	CVPR-20	47.7	54.1	-	-	-	-	72.3	72.9	-	71.0	71.7	-
	DDAG [50]	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 [51]	TPAMI-21	47.50	47.65	35.30	54.17	62.97	59.23	70.05	66.37	50.19	70.49	65.90	51.24
	VCD+VML [26]	CVPR-21	60.02	58.80	-	66.05	72.98	-	73.2	71.6	-	71.8	70.1	-
	CA [49]	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 [39]	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 [56]	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 [17]	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 [37]	ICCV-21	67.39	61.78	-	68.84	75.56	-	83.93	79.83	-	83.05	78.57	-
	FMCNet[54]	CVPR-22	66.34	62.51	-	68.15	74.09	-	89.12	84.43	-	88.38	83.86	-
	MAUM [22]	CVPR-22	71.68	68.79	-	76.97	81.94	-	87.87	85.09	-	86.95	84.34	-
Unsupervised	SSG [11]	ICCV-19	2.32	5.00	-	-	-	-	1.91	3.18	-	-	-	-
	ECN [59]	CVPR-19	8.07	12.68	-	-	-	-	2.17	2.90	-	-	-	-
	SPCL [15]	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 [14]	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 [40]	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 [32]	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 [7]	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 [2]	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 [5]	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 [31]	ECCV-22	29.9	27.1	-	29.8	38.8	-	32.9	29.7	-	32.1	28.6	-
	H2H [21]	TIP-21	30.15	29.40	-	-	-	-	23.81	18.87	-	-	-	-
	ADCA [43]	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
	GUR*(Ours)	-	60.95	56.99	41.85	64.22	69.49	64.81	**73.91**	**70.23**	**58.88**	**75.00**	**69.94**	**56.21**
	GUR (Ours)	-	**63.51**	**61.63**	**47.93**	**71.11**	**76.23**	**72.57**	-	-	-	-	-	-

Table 1. The comparison with the state-of-the-art methods on SYSU-MM01 and RegDB. It contains two groups, *i.e.*, unsupervised ReID methods and supervised VI-ReID methods. Rank at r accuracy(%), mAP (%) and mINP (%) are reported. GUR* denotes the results without camera information.

crepancy and enhancing the learning of modality-invariant features. With our insightful solutions, GUR achieves superior performance compared with existing unsupervised methods. In addition, the label distributions within each camera are unbalanced in the SYSU-MM01 dataset, *i.e.*, some cameras only contain part of the identities, increasing the difficulty of learning unified representation. The excellent performance demonstrates that our approach is also effective in learning from unbalanced label distribution data.

Comparison with Supervised Methods. Our GUR achieves competitive performance with VCD+VML [26], and even surpasses some supervised methods including Zero-Paddiing [38], eBDTR [48], HSME [18], AGW [51], DDAG[50], and so on. The excellent performance of our method benefits from the insightful design for the hierarchical discrepancy. There are three major advantages of our method: 1) Our learning framework is highly scalable and can be used in any contrastive learning with memory modules to handle domain gaps. 2) The learned features are robust to domain discrepancy at different levels. 3) Our method can also be utilized for other cross-modality retrieval tasks, *e.g.*, visible-infrared face recognition.

4.4. Ablation Study

To evaluate the contribution of each component, we conduct an ablation experiment on SYSU-MM01 and RegDB datasets, as shown in Table 2.

Baseline denotes the ADC [43] which adopts a dual-contrastive learning framework. Although ADC promotes unsupervised cross-modality learning, the hierarchical discrepancy hinders the further improvement of the discriminability of features for retrieval.

Effectiveness of Bottom-up Domain Learning Strategy. For the setting of only using the BD module, we remove the CAE module and use the original feature to cluster and assign pseudo labels. Compared with the baseline, bottom-up domain learning has a slight improvement in accuracy. This implies the difficulty of associating the same person from different domains without the CAE module. Indeed, when we only use BD, it provides a slight performance gain. But

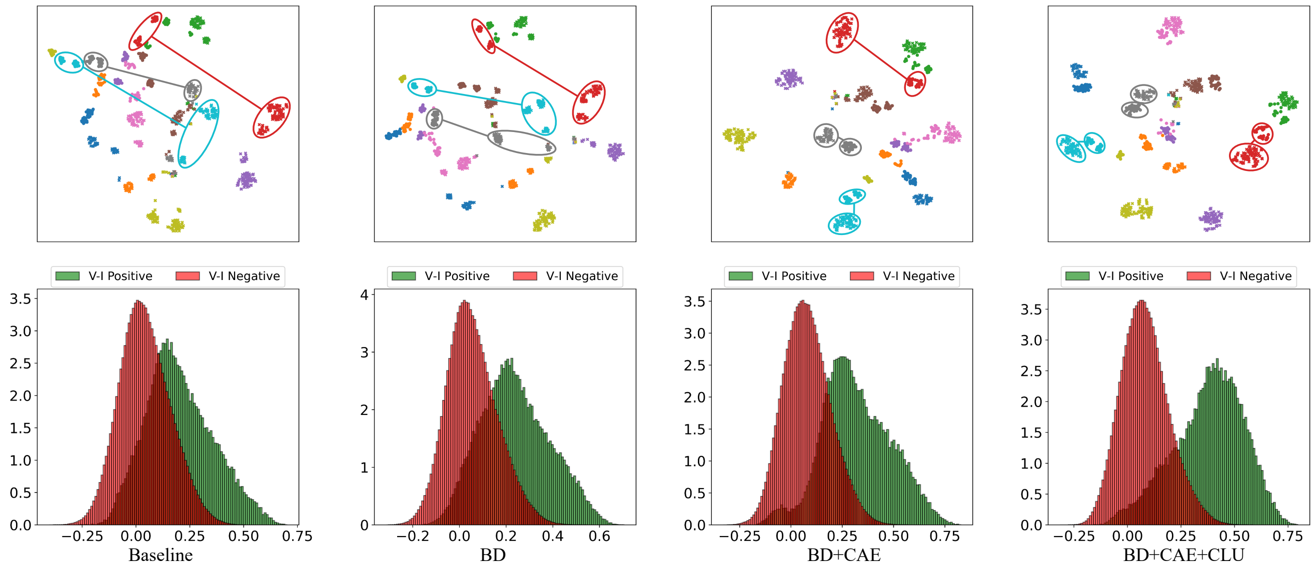

Figure 3. The t-SNE (first row) and similarity distribution (second row) visualization of 20 randomly selected identities. In t-SNE visualization, the color indicates the identity. Circle means visible modality and the cross means the infrared modality.

					SYSU-MM01						RegDB					
	Components				All Search			Indoor Search			Visible to Infrared			Infrared to Visible		
Index	Baseline	BD	CAE	CLU	$r1$	mAP	mINP	$r1$	mAP	mINP	$r1$	mAP	mINP	$r1$	mAP	mINP
1	✓				35.07	34.58	22.05	43.66	52.23	48.05	41.12	40.18	30.58	42.83	43.31	34.26
2		✓			36.63	36.70	24.43	42.00	51.12	47.39	43.42	42.51	32.74	43.88	41.78	30.85
3		✓	✓		55.96	55.62	42.79	62.93	70.43	66.75	69.13	68.54	61.67	68.84	67.69	58.49
4		✓		✓	54.27	52.30	38.17	60.01	66.78	63.02	59.74	59.84	52.76	64.67	63.00	53.52
5	✓		✓	✓	57.99	53.60	38.07	59.97	66.73	62.20	62.82	61.28	50.12	61.44	56.63	42.42
6		✓	✓	✓	**63.51**	**61.63**	**47.93**	**71.11**	**76.23**	**72.57**	**73.91**	**70.23**	**58.88**	**75.00**	**69.94**	**56.21**

Table 2. Ablation studies on the SYSU-MM01 and RegDB. "Baseline" means the augmented joint dual-contrastive learning framework [43]. "BD" represents the bottom-up domain learning strategy. Rank at r accuracy (%), mAP (%) and mINP (%) are reported.

when the BD is integrated with the CAE, GUR has significant improvement. The main reason is that BD and CAE are complementary and should be used in combination to handle hierarchical discrepancy. BD is a bottom-up domain learning strategy, *i.e.*, *intra-camera*, *inter-camera* and *inter-modality training*, and it optimizes the memory at different levels for CAE. Without BD, the memory is inaccurate for computing cross-memory association embedding.

Discussion of removing BD. We conduct the experiments of removing BD and only using CAE and CLU with baseline, as shown in the index 5 of Table 2 . Compared with the full GUR (BD+CAE+CLU), the performance of only using CAE and CLU has a drop of about 6%-10% rank-1 accuracy, which indicates that it is better to combine the BD and CAE to improve cross-modality retrieval. BD can provide better memory representations for CAE and CAE can bring reliable clustering for BD, formulating a mutual reinforcement.

Effectiveness of CAE. We observe significant improvement in accuracy when integrating the CAE module for bottom-up domain learning. The major advantage of CAE is that it can explore the relationship between images and each camera or modality memory and compose a unified representation embedding, which is a probability of association with memories and robust to camera and modality variations.

Effectiveness of CLU. Compared with the results in index 2 and 3, the experiments in index 4 and 5 demonstrate that the CLU module significantly improves the Rank-1 accuracy by about 10%-20%, greatly enhancing the cross-modality generalizability of learned features and ensuring the semantic consistency of the two modality labels.

Based on the above experiments, the CAE module and CLU module significantly improve the mAP and Rank-1 accuracy under various settings. These results demonstrate that each module plays a critical role in handling the hierarchical discrepancy in USL-VI-ReID.

4.5. Further Analysis

Hyper-parameter Analysis. The proposed GUR involves a key parameter σ in Eq 6. To study the effect of σ, we set it to different values as shown in Figure 4. The σ controls

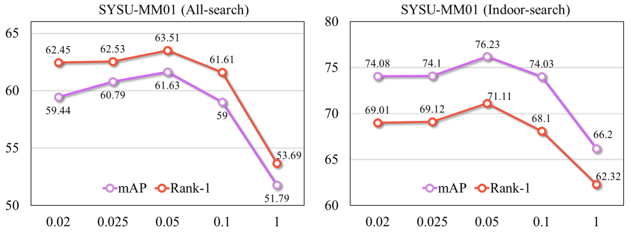

Figure 4. Evaluation of different σ in Eq 6. The results are based on all search mode (left) and indoor search mode (right) of SYSU-MM01 dataset. The Rank-1 (%) and mAP (%) are reported.

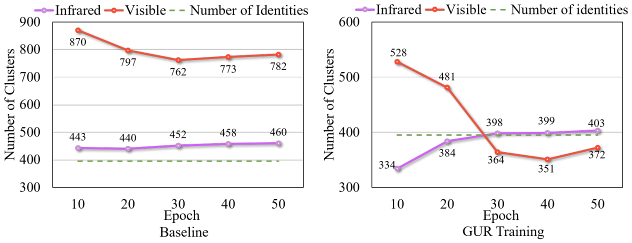

Figure 5. Evaluation of clustering consistencies. The results are based on all search mode of SYSU-MM01 dataset.

the concern on hard negative samples. Small σ penalizes much more on the hardest negative samples and has less tolerance to potential positive samples [29, 33]. Large σ makes the probability smooth and less sensitive to the hard negative samples. We find that the accuracy is significantly improved with a σ less than 1.0. When $\sigma = 0.05$, GUR achieves a balance in the dilemma of uniformity-tolerance and obtains the best results.

Clustering Consistencies. We aim to study the effect of our method on clustering consistencies, as shown in Fig. 5. We observe a large difference in the number of infrared and visible clusters in the baseline and the number of visible clusters is significantly less than the infrared clusters. The reason may be that visible camera is more susceptible to the light environment compared with infrared cameras, resulting in the split of identity. After GUR training, the numbers of infrared and visible clusters gradually approximate real identities, proving the effectiveness of our method for alleviating the clustering inconsistencies.

Visualization. We visualize the t-SNE [28] map and cosine similarity distribution of positive/negative cross-modality matching pairs of 20 randomly selected identities in Fig. 3. From 'Baseline' to 'BD+CAE+CLU', the sample points of the two modalities are gradually drawn closer and the infrared-visible positive/negative distributions are increasingly separated well. The above two visualizations demonstrate that our method results in better robustness of features against hierarchical discrepancy. We also note that some samples of the same identity are not clustered together, showing that there is still much room to improve for the USL-VI-ReID task.

5. Conclusion

This paper presents a grand unified representation (GUR) learning framework for USL-VI-ReID. We propose a bottom-up domain learning strategy with the cross-memory association embedding module to handle the hierarchical discrepancy, *i.e.* camera variation and modality discrepancy. Cross-memory association embedding module mines the relationship between images and each camera or modality memory and unifies the features across different cameras and modalities for reliable clustering. Furthermore, we introduce a cross-modality label unification module to optimize the generation of cross-modality labels, explicitly enhancing the centralization of positive cross-modality representations. Finally, with the above modules, GUR learns a unified and robust representation against the hierarchical discrepancy. Extensive experiments validate the superior performance of our method, further narrowing the gap between supervised and unsupervised visible-infrared person re-identification.

Limitations and Future Research. Although our approach achieves impressive performance, there are two limitations: 1) there is still much room to improve compared with supervised VI-ReID task. 2) our method needs more time for training, which can be improved. In the future, it is desirable to investigate the cross-modality label association with global and part features based on the transformer, which will be a useful method to refine the error labels and provide more accurate supervision for cross-modality learning.

Acknowledgments

This work is partially supported by the National Natural Science Foundation of China under Grants (62071338, 62176188, 61771180), the Key Research and Development Program of Hubei Province (2021BAA187, 2022BCA009), the Special Fund of Hubei Luojia Laboratory (220100015), Zhejiang Lab (NO.2022NF0AB01), and the CAAI-Huawei MindSpore Open Fund. The numerical calculations in this paper had been supported by the super-computing system in the Supercomputing Center of Wuhan University.

References

[1] Cuiqun Chen, Mang Ye, Meibin Qi, Jingjing Wu, Jianguo Jiang, and Chia-Wen Lin. Structure-aware positional transformer for visible-infrared person re-identification. *IEEE TIP*, pages 2352–2364, 2022.

[2] Hao Chen, Benoit Lagadec, and Francois Bremond. Ice: Inter-instance contrastive encoding for unsupervised person re-identification. In *ICCV*, pages 14960–14969, 2021.

[3] Hao Chen, Yaohui Wang, Benoit Lagadec, Antitza Dantcheva, and Francois Bremond. Joint generative and contrastive learning for unsupervised person re-identification. In *CVPR*, pages 2004–2013, 2021.

[4] Yanbei Chen, Xiatian Zhu, and Shaogang Gong. Instance-guided context rendering for cross-domain person re-identification. In *ICCV*, pages 232–242, 2019.

[5] Yoonki Cho, Woo Jae Kim, Seunghoon Hong, and Sung-Eui Yoon. Part-based pseudo label refinement for unsupervised person re-identification. In *CVPR*, pages 7308–7318, 2022.

[6] Seokeon Choi, Sumin Lee, Youngeun Kim, Taekyung Kim, and Changick Kim. Hi-cmd: Hierarchical cross-modality disentanglement for visible-infrared person re-identification. In *CVPR*, pages 10257–10266, 2020.

[7] 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.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

[9] Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. *NIPS*, 32, 2019.

[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*, pages 226–231, 1996.

[11] Yang Fu, Yunchao Wei, Guanshuo Wang, Yuqian Zhou, Honghui Shi, and Thomas S Huang. Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In *ICCV*, pages 6112–6121, 2019.

[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 *Proceedings of the 29th ACM International Conference on Multimedia*, pages 5257–5265, 2021.

[13] 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 *ACM MM*, pages 3644–3653, 2021.

[14] 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.

[15] Yixiao Ge, Feng Zhu, Dapeng Chen, Rui Zhao, et al. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. *NeurIPS*, pages 11309–11321, 2020.

[16] Jianyuan Guo, Yuhui Yuan, Lang Huang, Chao Zhang, Jin-Ge Yao, and Kai Han. Beyond human parts: Dual part-aligned representations for person re-identification. In *ICCV*, pages 3642–3651, 2019.

[17] Xin Hao, Sanyuan Zhao, Mang Ye, and Jianbing Shen. Cross-modality person re-identification via modality confusion and center aggregation. In *ICCV*, pages 16403–16412, 2021.

[18] Yi Hao, Nannan Wang, Jie Li, and Xinbo Gao. Hsme: Hypersphere manifold embedding for visible thermal person re-identification. In *AAAI*, pages 8385–8392, 2019.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[20] Diangang Li, Xing Wei, Xiaopeng Hong, and Yihong Gong. Infrared-visible cross-modal person re-identification with an x modality. In *AAAI*, pages 4610–4617, 2020.

[21] Wenqi Liang, Guangcong Wang, Jianhuang Lai, and Xiaohua Xie. Homogeneous-to-heterogeneous: Unsupervised learning for rgb-infrared person re-identification. *IEEE TIP*, pages 6392–6407, 2021.

[22] 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*, pages 19366–19375, 2022.

[23] 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*, pages 13379–13389, 2020.

[24] 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*, page 605, 2017.

[25] Liangchen Song, Cheng Wang, Lefei Zhang, Bo Du, Qian Zhang, Chang Huang, and Xinggang Wang. Unsupervised domain adaptive re-identification: Theory and practice. *Pattern Recognition*, page 107173, 2020.

[26] Xudong Tian, Zhizhong Zhang, Shaohui Lin, Yanyun Qu, Yuan Xie, and Lizhuang Ma. Farewell to mutual information: Variational distillation for cross-modal person re-identification. In *CVPR*, pages 1522–1531, 2021.

[27] Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. Simultaneous deep transfer across domains and tasks. In *ICCV*, pages 4068–4076, 2015.

[28] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[29] Feng Wang and Huaping Liu. Understanding the behaviour of contrastive loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2495–2504, 2021.

[30] Guan'an Wang, Tianzhu Zhang, Jian Cheng, Si Liu, Yang Yang, and Zengguang Hou. Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment. In *ICCV*, pages 3623–3632, 2019.

[31] 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. 2022.

[32] Menglin Wang, Baisheng Lai, Jianqiang Huang, Xiaojin Gong, and Xian-Sheng Hua. Camera-aware proxies for unsupervised person re-identification. In *AAAI*, page 4, 2021.

[33] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR, 2020.

[34] 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*, pages 618–626, 2019.

[35] 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.

[36] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, pages 79–88, 2018.

[37] Ziyu Wei, Xi Yang, Nannan Wang, and Xinbo Gao. Syncretic modality collaborative learning for visible infrared person re-identification. In *ICCV*, pages 225–234, 2021.

[38] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. Rgb-infrared cross-modality person re-identification. In *ICCV*, pages 5380–5389, 2017.

[39] 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*, pages 4330–4339, 2021.

[40] Shiyu Xuan and Shiliang Zhang. Intra-inter camera similarity for unsupervised person re-identification. In *CVPR*, pages 11926–11935, 2021.

[41] Shiyu Xuan and Shiliang Zhang. Intra-inter domain similarity for unsupervised person re-identification. *IEEE TPAMI*, 2022.

[42] Bin Yang, Jun Chen, and Mang Ye. Top-k visual tokens transformer: Selecting tokens for visible-infrared person re-identification. In *ICASSP*, pages 1–5. IEEE, 2023.

[43] Bin Yang, Mang Ye, Jun Chen, and Zesen Wu. Augmented dual-contrastive aggregation learning for unsupervised visible-infrared person re-identification. In *ACM MM*, page 2843–2851, 2022.

[44] Mang Ye, Cuiqun Chen, Jianbing Shen, and Ling Shao. Dynamic tri-level relation mining with attentive graph for visible infrared re-identification. *IEEE TIFS*, 17:386–398, 2021.

[45] Mang Ye, Xiangyuan Lan, Qingming Leng, and Jianbing Shen. Cross-modality person re-identification via modality-aware collaborative ensemble learning. *IEEE TIP*, pages 9387–9399, 2020.

[46] Mang Ye, Xiangyuan Lan, Jiawei Li, and Pong Yuen. Hierarchical discriminative learning for visible thermal person re-identification. In *AAAI*, page 7501–7508, 2018.

[47] Mang Ye, Xiangyuan Lan, Jiawei Li, and Pong Yuen. Hierarchical discriminative learning for visible thermal person re-identification. In *AAAI*, 2018.

[48] 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.

[49] Mang Ye, Weijian Ruan, Bo Du, and Mike Zheng Shou. Channel augmented joint learning for visible-infrared recognition. In *ICCV*, pages 13567–13576, 2021.

[50] 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*, pages 229–247, 2020.

[51] 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.

[52] Mang Ye, Jianbing Shen, and Ling Shao. Visible-infrared person re-identification via homogeneous augmented tri-modal learning. *IEEE TIFS*, pages 728–739, 2020.

[53] Hong-Xing Yu, Ancong Wu, and Wei-Shi Zheng. Unsupervised person re-identification by deep asymmetric metric embedding. *IEEE TPAMI*, pages 956–973, 2020.

[54] Qiang Zhang, Changzhou Lai, Jianan Liu, Nianchang Huang, and Jungong Han. Fmcnet: Feature-level modality compensation for visible-infrared person re-identification. In *CVPR*, pages 7349–7358, 2022.

[55] Yukang Zhang and Hanzi Wang. Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification. In *CVPR*, pages 2153–2162, 2023.

[56] Yukang Zhang, Yan Yan, Yang Lu, and Hanzi Wang. Towards a unified middle modality learning for visible-infrared person re-identification. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 788–796, 2021.

[57] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, pages 1116–1124, 2015.

[58] Zhun Zhong, Liang Zheng, Shaozi Li, and Yi Yang. Generalizing a person retrieval model hetero-and homogeneously. In *ECCV*, pages 172–188, 2018.

[59] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *CVPR*, pages 598–607, 2019.

[60] Yang Zou, Xiaodong Yang, Zhiding Yu, BVK Kumar, and Jan Kautz. Joint disentangling and adaptation for cross-domain person re-identification. In *ECCV*, pages 87–104, 2020.