Unsupervised Visible-Infrared Person Re-Identification via Progressive Graph Matching and Alternate Learning

Zesen Wu[1], Mang Ye[1,2*]
[1]National Engineering Research Center for Multimedia Software, Institute of Artificial Intelligence,
Hubei Key Laboratory of Multimedia and Network Communication Engineering,
School of Computer Science, Wuhan University, Wuhan, China
[2] Hubei Luojia Laboratory, Wuhan, China

https://github.com/zesenwu23/USL-VI-ReID

Abstract

Unsupervised visible-infrared person re-identification is a challenging task due to the large modality gap and the unavailability of cross-modality correspondences. Cross-modality correspondences are very crucial to bridge the modality gap. Some existing works try to mine cross-modality correspondences, but they focus only on local information. They do not fully exploit the global relationship across identities, thus limiting the quality of the mined correspondences. Worse still, the number of clusters of the two modalities is often inconsistent, exacerbating the unreliability of the generated correspondences. In response, we devise a Progressive Graph Matching method to globally mine cross-modality correspondences under cluster imbalance scenarios. PGM formulates correspondence mining as a graph matching process and considers the global information by minimizing the global matching cost, where the matching cost measures the dissimilarity of clusters. Besides, PGM adopts a progressive strategy to address the imbalance issue with multiple dynamic matching processes. Based on PGM, we design an Alternate Cross Contrastive Learning (ACCL) module to reduce the modality gap with the mined cross-modality correspondences, while mitigating the effect of noise in correspondences through an alternate scheme. Extensive experiments demonstrate the reliability of the generated correspondences and the effectiveness of our method.

1. Introduction

The target of visible-infrared person re-identification (VI-ReID) [23, 25, 38, 51, 52] is to recognize the same person across a set of visible/infrared gallery images when given an image from another modality. This task has attracted extensive interest recently due to its significance in night intelligent surveillance and public security. Many progresses [3, 5, 29, 40, 51] have been made in VI-ReID. However, these methods require well-annotated training sets which are exhausting to obtain, so they are less applicable

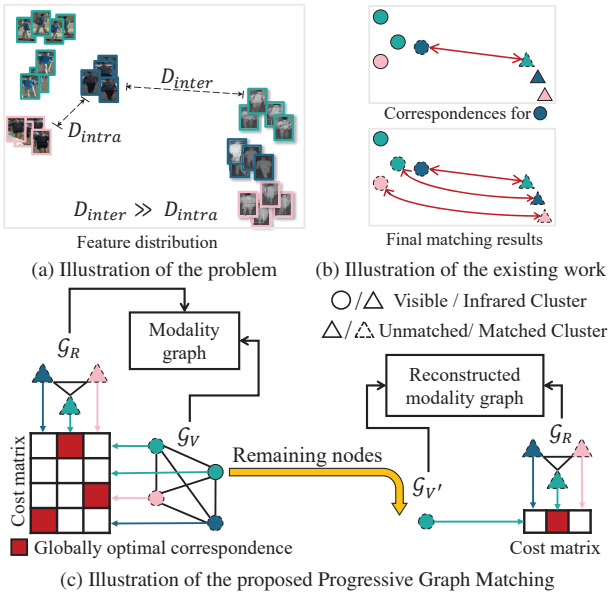

Figure 1. Idea illustration. Different colors indicate different pedestrians. (a) illustrates the feature distribution of randomly selected persons of SYSU-MM01. The cross-modality discrepancy is much larger than inter-class variance within each modality. (b) abstracts the existing solution. The locally closest unmatched cross-modality cluster is treated as correspondence. Bottom of (b) indicates its two drawbacks: 1) it ignores the global information among different identities and 2) it ignores the cluster imbalance issue across modalities and discards remaining nodes (●). (c) is the progressive graph matching method. We utilize graph matching to obtain the globally optimal correspondences and design a progressive strategy to handle the cluster imbalance issue.

*Corresponding Author: Mang Ye

in real scenarios. In light of this limitation, we attempt to investigate an unsupervised solution for VI-ReID.

For unsupervised single-modality ReID, widely-studied works [4, 7, 9, 34, 42, 57] utilize cluster-based methods to produce supervision signals in the homogeneous space. However, in the visible-infrared heterogeneous space, the consistency of features and semantics cannot be maintained due to the large modality gap. Specifically, the cross-modality difference is much larger than the inter-class discrepancy within each modality (see Fig. 1a). Hence, we cannot establish connections between the two modalities by adopting off-the-shelf clustering methods. However, cross-modality correspondences play an important role in bridging the modality gap between two heterogeneous modalities [25, 29, 40, 51, 52]. Without reliable cross-modality correspondences, the model can hardly learn modality-invariant features.

Some efforts [22, 33, 45] have been made recently to find cross-modality correspondences. However, most of the existing methods consider only local information and do not take full advantage of the global relationship among different identities (see Fig. 1b). What's worse, they are not applicable to scenarios with cluster imbalance problems, since some clusters cannot find their correspondences, hindering the later modality gap reduction process. To globally mine cross-modality correspondences under cluster imbalance scenarios, we propose a Progressive Graph Matching (PGM) method. It is featured for two designs, *i.e.*, 1) connecting the two modalities with graph matching and 2) addressing the imbalance issue with the progressive strategy.

First, we employ graph matching to fully utilize the relationship among different identities under global constraints (see Fig. 1c left). PGM formulates the cross-modality correspondences mining process as a bipartite graph matching problem with each modality as a graph and each cluster as a node. The matching cost between nodes is positively correlated with the distance of clusters. By minimizing the global matching cost, graph matching is expected to generate more reliable correspondences with global consideration. Graph matching has been demonstrated to have an advantage in unsupervised correspondence localization between two feature sets [6, 35, 44, 49, 50]. With this property, we are inspired to construct a graph for each modality and link the same person across different modalities.

Second, we propose a progressive strategy to tackle the imbalance problem. Basic graph matching cannot handle the cluster imbalance issue across modalities, which is caused by camera variations within class. Instances of the same person are sometimes split into different clusters [4, 57] and some clusters cannot find their cross-modality correspondences (see Fig. 1c). This correspondence-missing problem affects the further modality discrepancy decrease. In response, we propose to find the correspondence for each cluster through multiple dynamic matching (see Fig. 1c right). The subgraphs in the bipartite graph are dynamically updated according to the previous matching results until each cluster progressively finds its correspondence. With the progressive strategy, different clusters with the same person ID could find the same cross-modality correspondences. Therefore, these many-to-one matching results alleviate the imbalance issue and also implicitly enhance intra-class compactness.

In addition, to fully exploit the mined cross-modality correspondences, we design a novel Alternate Cross Contrastive Learning (ACCL) module. Inspired by supervised methods like [23, 25, 47], Cross Contrastive Learning (CCL) reduces the modality discrepancy by pulling the instance close to its corresponding cross-modality proxy and pushing it away from other proxies. However, unlike the supervised setting, the cross-modality correspondences generated by unsupervised methods are inevitably noisy, so directly combining the two unidirectional metric losses (*visible to infrared* and *infrared to visible*) may lead to rapid false "association". We propose to alternately use two unidirectional metric losses so that positive cross-modality pairs can be associated by stages. This alternate scheme mitigates the effect of noise since the false positive pairs do not remain for long. In an alternative way, the noise effect would be reduced (as detailed in Sec. 3.3).

Our main contributions can be summarized as follows:

- We propose the PGM method to mine reliable cross-modality correspondences for unsupervised VI-ReID. We first build modality graph and perform graph matching to consider global information among identities and devise a progressive strategy to make the matching process applicable to imbalanced clusters.
- We design ACCL to decrease the modality disparity, which promotes the learning of modality-invariant information by gathering the instance to its corresponding cross-modality proxy. The alternate updating scheme is designed to mitigate the effect of noisy cross-modality correspondences.
- Extensive experiments demonstrate that PGM method provides relatively reliable cross-modality correspondences and our proposed method achieves significant improvement in unsupervised VI-ReID.

2. Related Work

2.1. Visible-Infrared Person ReID

Supervised VI-ReID has drawn increasing interest recently due to its potential in 24-hour surveillance. It mainly suffers from the modality discrepancy originating from different spectrum cameras [38]. To alleviate the cross-modality discrepancy, many works apply feature-level constraints to embed heterogeneous images into a shared fea-

ture space so as to align feature distribution [23, 25, 40, 47]. Among these, [25] utilizes a unidirectional cross-modality metric to alleviate the relay effect and promote modality association. Another representative method-of-choice is to make up the missing modality-specific information from existing modalities [21, 32, 36, 55, 58]. Zhang first *et al*. proposed an FMCNet [55] to compensate for missing modality-specific information at the feature level rather than the image level. However, the success of the above-described supervised approaches is partially attributable to the availability of well-annotated training datasets.

Unsupervised VI-ReID is raised to cope with the lack of annotations. H2H [22] makes the first attempt to address this challenging problem by proposing a two-stage learning approach. In OTLA [33], Wang *et al*. try to assign the infrared images to the pseudo visible labels based on the optimal-transport strategy. These methods require extra RGB datasets for pre-training and OTLA also assumes each visible label is assigned to a similar number of infrared images, which may not hold in practice. Yang *et al*. first mine the cluster-level relationship [45] with cross-modality memory aggregation but it lacks global consideration and cannot handle the cluster imbalance issue.

2.2. Unsupervised Person ReID

In an effort to alleviate the conflict between annotation and performance, unsupervised ReID has attracted increasing attention. These methods can be roughly classified as Unsupervised Domain Adaption (UDA) and UnSupervised Learning (USL) methods. The target of UDA-based methods is to adapt models trained on labeled source domain to unlabeled target domain [28]. Among UDA-based methods, several works [19, 26, 63, 64] attempt to reduce domain gap by finding positive or negative pairs from labeled source and unlabeled target dataset. Some [11, 37, 62] would like to employ generative networks to transfer images of source domain into the style of target domain. Another possibility is to acquire pseudo labels by clustering methods from target domain [1, 13–15, 60]. The USL methods [7, 24, 31, 43, 46, 56, 57, 61] are mainly based on pseudo labels, which establish a bridge with supervised manner. However, due to the large modal discrepancies between visible and infrared images, the unsupervised methods designed for single-modality ReID are not applicable for visible-infrared ReID.

2.3. Graph Matching for Person Re-ID

In the context of single-modality ReID, graph matching is mainly utilized in two ways. **1)** The pedestrian image is divided into slices or parts, and each slice or part is considered as a node inside the graph [41, 59]. Graph matching is used to align parts of different person images. **2)** In [16, 39, 50], each camera view is considered as a graph and each person within the camera is considered as a node. Graph matching is used to identify the same person across multiple cameras. For VI-ReID, however, the cross-modality difference is much larger than inter-camera variance within each modality, thus we construct a graph for each modality and explore the correspondences across modalities with graph matching.

3. Methodology

The framework of our proposed method is illustrated in Fig. 2. We first utilize the Dual-Contrastive Learning (DCL [45]) framework to learn intra-modality discriminability, which is optimized by the joint intra-modality contrastive learning. Based on DCL, the proposed method lays emphasis on its novel progressive graph matching (middle in Fig. 2) and alternate cross contrastive learning module (right in Fig. 2), which are described in detail in Sec. 3.2 and Sec. 3.3, respectively.

3.1. Dual-Contrastive Learning framework

Given a visible-infrared training dataset $\mathcal{T} = \{\mathcal{T}^v, \mathcal{T}^r\}$, $\mathcal{T}^v = \{x_i^v | i = 1, 2, \cdots, N\}$ represents visible dataset with N visible instances and $\mathcal{T}^r = \{x_i^r | i = 1, 2, \cdots, M\}$ denotes M infrared images. It should be noted that channel augmentation [51] is a common and powerful data augmentation to bridge the gap between visible and infrared images, and thus channel augmented (CA) images are used to assist in the learning process of visible streams.

The two-stream backbone (*e.g*., ResNet50 [17] and AGW [53]) f is used to extract the features of these pedestrian images. Visible and infrared memories are constructed after their features got clustered by DBSCAN [12]. $\mathcal{K}^e \in \mathbb{R}^{d \times Y^e}$ is the memory for modality e ($e = \{v, r\}$, indicating visible and infrared modality, respectively), where d is the feature dimension and Y^e is the number of clusters for modality e. Each proxy represents all the instances of the same cluster and each entry of the memory is initialized with the mean feature of its corresponding proxy. The memory is updated by

$$\mathcal{K}^e[j] \leftarrow \lambda \mathcal{K}^e[j] + (1 - \lambda) f(x_i^e), \tag{1}$$

where $\mathcal{K}^e[j]$ stores the feature centroid for $j\text{-}th$ class in modality e. Besides, x_i^e is an image in class j and $\lambda \in [0, 1]$ is the memory updating rate.

In a mini-batch during training, we randomly sample P classes and K samples per class as in [18] for the infrared modality. Considering that each visible image has its augmented CA image, to balance the number of images of different modalities, we randomly choose P classes, each containing $(K/2)$ visible images and their generated $(K/2)$ CA images. For infrared modality, a ClusterNCE [9] loss is:

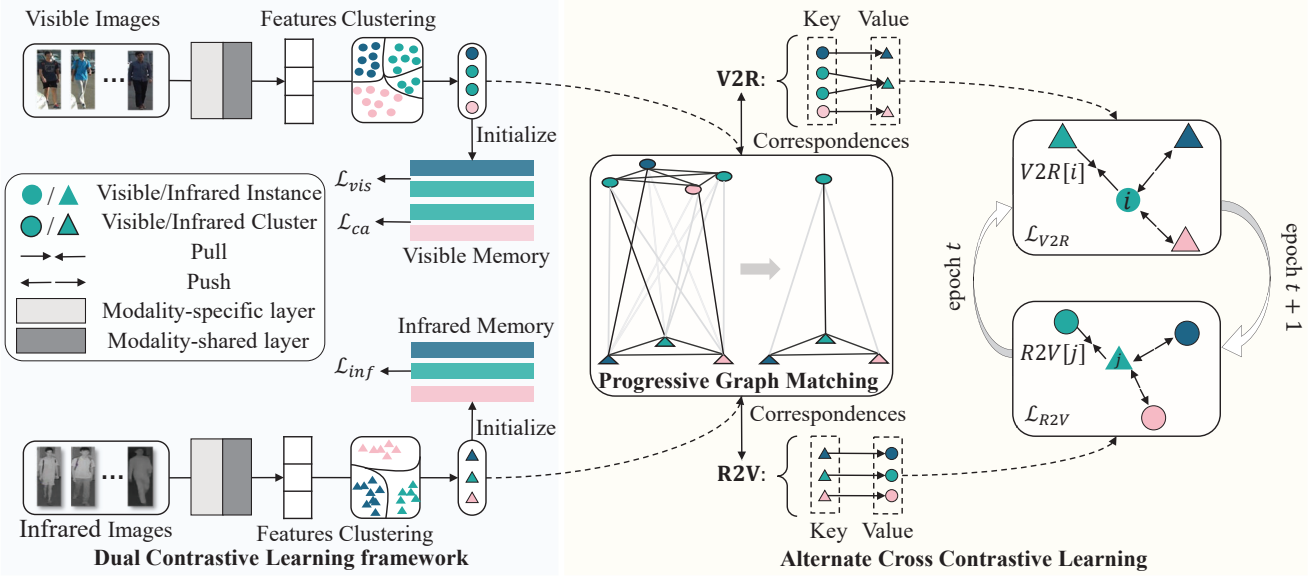

Figure 2. The pipeline of our framework. Different colors indicate different pedestrians. It contains the Dual Contrastive Learning framework (baseline, described in Sec. 3.1) and two key novel components: Progressive Graph Matching (PGM, described in Sec. 3.2) method and Alternate Cross Contrastive Learning (ACCL, described in Sec. 3.3) module. PGM method is proposed to find reliable cross-modality correspondences (stored in $\mathbf{V2R}$ and $\mathbf{R2V}$), which take part in the further ACCL to learn modality-invariant features.

$$\mathcal{L}_{inf} = -\sum_{i=1}^{N_B} \log\left(\frac{\exp(\mathcal{K}^r[\widetilde{y_i^r}]^T \cdot f(x_i^r)/\tau)}{\sum_{k=1}^{Y^r} \exp(\mathcal{K}^r[k]^T \cdot f(x_i^r)/\tau)}\right), \quad (2)$$

where $N_B = P \times K$ is the number of infrared instances and $\widetilde{y_i^r}$ is the class (pseudo label) for image x_i^r. Besides, τ is a temperature factor. This loss achieves classification by gathering an instance toward the proxy of its class while scattering all other proxies. The loss function for the visible modality and its CA modality is denoted by $\mathcal{L}_{vis}$ and $\mathcal{L}_{ca}$, respectively. Their formulations are similar to Eq. (2) and are omitted here. The details of $\mathcal{L}_{vis}$ and $\mathcal{L}_{ca}$ are given in the supplementary material.

$$\mathcal{L}_{dcl} = \mathcal{L}_{inf} + \mathcal{L}_{vis} + \mathcal{L}_{ca} \quad (3)$$

The DCL loss function $\mathcal{L}_{dcl}$ combines these ClusterNCE [9] losses, which help the model to learn intra-modality discriminability and the augmented stream assists to learn certain modality-invariant features.

3.2. Progressive Graph Matching

DCL mentioned above does not directly explore the relation between the two modalities and thus can not cope with the case of excessive modality differences. To connect visible and infrared data, we present the PGM method to find reliable cross-modality correspondences.

Notaion Definition. We build a graph for each modality and each graph can be viewed as part of a bipartite graph. Suppose that the visible graph $\mathcal{G}_\mathcal{V}$ contains Y^v nodes (clusters), which can be denoted by $[\mathcal{V}] = \{c_i^v | i = 1, 2, \cdots, Y^v\}$. Analogously, the infrared graph $\mathcal{G}_\mathcal{R}$ includes Y^r infrared nodes represented by $[\mathcal{R}] = \{c_j^r | j = 1, 2, \cdots, Y^r\}$. We use $C = \{C(i,j)\}$ to denote the assignment cost matrix with each element illustrating the dissimilarity of node c_i^v and node c_j^r. Our target is to find the correspondence in $\mathcal{G}_\mathcal{R}$ ($\mathcal{G}_\mathcal{V}$) for each node in $\mathcal{G}_\mathcal{V}$ ($\mathcal{G}_\mathcal{R}$). We assume that $Y^v \geq Y^r$, indicating that the number of clusters in the two modalities is different.

Cost Matrix. The assignment cost in the graph matching method can be represented by the dissimilarity between the features of different clusters under a certain metric. The basic idea of the matching cost is to penalize the matched clusters across two modalities with the feature difference. That is, the more similar the features of clusters are, the lower the cost is. We design a simple yet effective cost expression, which is formulated by

$$\begin{aligned} C(i,j) &= \frac{1}{\exp(Sim(i,j))}, \\ Sim(i,j) &= \frac{u_i^v \cdot u_j^r}{||u_i^v|| \times ||u_j^r||}, \\ u_i^v &= \frac{1}{|c_i^v|} \sum_{x_i^v \in c_i^v} f(x_i^v), \\ u_j^r &= \frac{1}{|c_j^r|} \sum_{x_j^r \in c_j^r} f(x_j^r), \end{aligned} \quad (4)$$

where $|c_i^v|$ denotes the number of instances in the cluster c_i^v

and x_i^v (x_j^r) is the instance within the cluster c_i^v (c_j^r). The mean feature of instances inside the cluster is expressed as the cluster's representation.

Basic Graph Matching Formulation. We give a definition of Basic Graph Matching (BGM) formulation following [30], which can be formulated as binary linear programming with linear constraints:

$$\begin{aligned} G(\mathbf{m}) &= \arg\min_{\mathbf{m}} C^T \mathbf{m} \\ s.t.\ \forall i \in [\mathcal{V}], \forall j \in [\mathcal{R}] &: m_i^j \in \{0,1\}, \\ \forall i \in [\mathcal{V}] &: \sum_{j \in [\mathcal{R}]} m_i^j \leq 1, \\ \forall j \in [\mathcal{R}] &: \sum_{i \in [\mathcal{V}]} m_i^j = 1, \end{aligned} \tag{5}$$

where $\mathbf{m} = \{m_i^j\} \in \mathbb{R}^{Y^v * Y^r \times 1}$ is an indicator of the matching of nodes c_i^v and c_j^r, indicating whether c_i^v and c_j^r belong to the same person ($m_i^j = 1$) or not ($m_i^j = 0$). Various efficient solutions such as the Hungarian algorithm [2] could be used to solve the basic matching problem, so we will not describe these algorithms in detail.

Given cost matrix C, the BGM outputs matrix $\mathbf{m}$, which has Y^r elements of 1, representing Y^r matched positive pairs. Note that **not** each node c_i^v in $\mathcal{G}_\mathcal{V}$ can find a node c_j^r that satisfies $m_i^j = 1$ (see 3^{rd} line in Eq. (5)), indicating that there exit some clusters in visible modality not finding their correspondences. Further, we will introduce the PGM method to handle the imbalanced problem during matching.

Progressive Graph Matching Method. The core idea of the PGM method is to find the correspondence for each node through multiple dynamic matching. Specifically, we suppose that the nodes in $\mathcal{G}_\mathcal{V}$ are more than nodes in $\mathcal{G}_\mathcal{R}$. After performing one BGM process, there are remaining nodes in $\mathcal{G}_\mathcal{V}$ not finding their correspondences, and nodes in $\mathcal{G}_\mathcal{R}$ all find their correspondences. We dynamically reconstruct a new graph (annotated as $\mathcal{G}_{\mathcal{V}'}$) with the remaining nodes in $\mathcal{G}_\mathcal{V}$ and the edges between them. $\mathcal{G}_{\mathcal{V}'}$ and $\mathcal{G}_\mathcal{R}$ are recomposed into a bipartite graph and a new BGM process will be performed. Note that nodes in $\mathcal{G}_\mathcal{R}$ would not update their correspondences since they have already found one. In the new BGM process, only nodes in $\mathcal{G}_{\mathcal{V}'}$ will update their correspondences. The BGM is performed repeatedly until each node finds its correspondence progressively. Details are presented in Algorithm 1.

Algorithm 1: Progressive Graph Matching

Input: The cost matrix $C \in \mathbb{R}^{Y^v \times Y^r}$ (Suppose $Y^v \geq Y^r$).
Output: The two mapping dictionaries: $\mathbf{V2R}$ and $\mathbf{R2V}$, with keys storing clusters and values storing their correspondences.

Initialize and empty three tag arrays: $\mathbf{matched\_v}$, $\mathbf{unmatched\_v}$ and $\mathbf{matched\_r}$;
while len($\mathbf{matched\_v}$) $\neq Y^v$ **do**
  // basic graph matching result
  $\mathbf{m}$ = BGM (C) // through Eq. (5)
  for *each* $m_i^j \in \mathbf{m}$ **do**
    if $m_i^j = 1$ **then**
      if $i \notin \mathbf{matched\_v}$ **then**
        $\mathbf{V2R}[i] = j$;
        $\mathbf{matched\_v}$.append(i);
      if $j \notin \mathbf{matched\_r}$ **then**
        $\mathbf{R2V}[j] = i$;
        $\mathbf{matched\_r}$.append(j);
    else
      // remaining visible nodes
      $\mathbf{unmatched\_v}$.append(i);
  // update C with remaining nodes
  $C \leftarrow C[\mathbf{unmatched\_v}]$;
return $\mathbf{V2R}$, $\mathbf{R2V}$;

3.3. Alternate Cross Contrastive Learning

With the cross-modality correspondences obtained by the PGM method, we propose the ACCL to reduce the modality discrepancy while mitigating the effect of the noise in the correspondences.

Cross Contrastive Learning (CCL). The CCL consists of two unidirectional learning, namely *infrared to visible* (*R2V*) learning and *visible to infrared* (*V2R*) learning. The former can be expressed as:

$$\mathcal{L}_{R2V} = -\sum_{i=1}^{N_B} \log\left(\frac{\exp(\mathcal{K}^v[\hat{y_i^r}]^T \cdot f(x_i^r)/\tau)}{\sum_{k=1}^{Y^v} \exp(\mathcal{K}^v[k]^T \cdot f(x_i^r)/\tau)}\right), \tag{6}$$

where $\hat{y_i^r} = \mathbf{R2V}[\widetilde{y_i^r}]$. $\widetilde{y_i^r}$ is the pseudo label for the infrared image x_i^r, and $\hat{y_i^r}$ is the cross-modality correspondence for $\widetilde{y_i^r}$, also the cross-modality label for x_i^r. The CCL can bridge modality gap by gathering the given sample to its corresponding cross-modality proxy. *Visible to infrared* learning exhibits a similar form with the addition of CA assisted learning, denoted as $\mathcal{L}_{V2R}$. The difference is that half of the N_B images are visible images and the other half are their corresponding CA images. The details of $\mathcal{L}_{V2R}$ are given in the supplementary material.

Alternate CCL (ACCL). An intuitive loss function for CCL is to combine $\mathcal{L}_{V2R}$ and $\mathcal{L}_{R2V}$, expressed as

$$\mathcal{L}_{ccl} = \mathcal{L}_{V2R} + \mathcal{L}_{R2V}. \tag{7}$$

However, this combination would amplify the noise in cross-modality correspondences, leading to a false association of false positive pairs. We devise an alternate updating

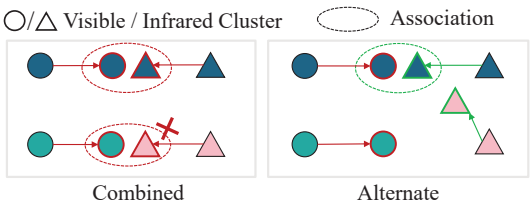

Figure 3. Comparison of the combined scheme ($\mathcal{L}_{ccl}$) and alternate ($\mathcal{L}_{accl}$) scheme. Color in ○ and △ indicates person ID. ← and ← indicate the pulling force in the epoch t and epoch $t+1$, respectively. Red and green edge represent the positions of the cluster at the epoch t and epoch $t+1$. In the epoch t, $\mathcal{L}_{ccl}$ associates the two pairs (true or false) by strong bidirectional force. For $\mathcal{L}_{accl}$, the true pair is associated after two rounds of unidirectional forces, while the false pair cannot be associated if they are not matched in the next epoch.

scheme where the cross-modality learning chooses a unidirectional metric in different iteration epochs, which can be expressed as:

$$\mathcal{L}_{accl} = \begin{cases} \mathcal{L}_{V2R}, & epoch \,\%\, 2 = 0 \\ \mathcal{L}_{R2V}, & epoch \,\%\, 2 = 1, \end{cases} \tag{8}$$

where $epoch$ indicates the index of iterations.

The overall loss is formulated as a combination of $\mathcal{L}_{dcl}$ and the cross-modality contrastive learning loss with a weighting parameter μ:

$$\mathcal{L}_{all} = \mathcal{L}_{dcl} + \mu \mathcal{L}_{accl}. \tag{9}$$

Rationale Analysis. We alternately utilize two unidirectional metric losses ($\mathcal{L}_{accl}$) instead of directly combining them as a bidirectional loss ($\mathcal{L}_{ccl}$). The directly combined bidirectional loss is commonly used in the supervised setting [25, 48], serving to associate the two modalities. However, it is not applicable in unsupervised scenarios. The reason is that the cross-modality correspondences obtained by unsupervised methods are inevitably unstable and noisy, and the "strong" bidirectional force will amplify the noise and lead to a false association. The alternate unidirectional learning acts to associate the two different modalities and mitigate the effect of noise with two novel designs, *i.e.*, **1)** learning unidirectional metrics and **2)** alternating unidirectional metrics. The former operates as a "weaker" force to prevent the association of false pairs too rapidly, compared to the bidirectional metric. The latter assures that the model is not biased towards a certain modality and that the true pairs can be progressively associated through multiple alternations. False pairs typically do not remain for long, therefore they cannot be progressively associated (see Fig. 3).

4. Experiments

4.1. Datasets and Evaluation Protocols

Datasets. We evaluate the proposed method on two widely used visible-infrared datasets SYSU-MM01 [38] and RegDB [27]. SYSU-MM01 contains 22,257 visible and 11,909 near-infrared images from 4 visible cameras and 2 infrared cameras in both indoor and outdoor environments. RegDB is a smaller and thus less challenging dataset that is collected by two aligned cameras (one visible and one infrared), and it consists of 412 person identities, where each identity has 10 visible images and 10 infrared images.

Evaluation Metrics. We follow the commonly used protocols [52] to evaluate both two datasets, where cumulative matching characteristic (CMC), mean average precision (mAP) and mean Inverse Negative Penalty (mINP [53]) are adopted. On SYSU-MM01, there are two different testing settings (*all-search* and *indoor-search* modes). The gallery is composed of visible images and the query consists of infrared images in the *all-search* mode. For *indoor-search* mode, images captured by visible outdoor scenes (CAM4 and CAM5) are discarded. RegDB contains two testing settings, including *thermal to visible* and *visible to thermal* modes. Following [51], we randomly split the training and testing set 10 times and report the overall average result.

Implementation Details. We adopt a non-local module enhanced network following AGW [53], which utilizes ResNet50 [17] as the feature extractor. The backbone parameters are initialized with the ImageNet [10] pre-trained weights. In a mini-batch, the number of classes P and samples for each class K are both 16. All the pedestrian images are resized to 288 × 144. We use Adam optimizer to train the model with weight decay 5e-4. The initial learning rate is 3.5×10^{-3} and decays 10 times every 20 epochs. The DCL network is trained for 50 epochs. Then we train the whole network with the pre-trained DCL for another 50 epochs. The augmentations for visible and visible images are following [45]. Besides, visible images are also augmented with Random GrayScale, which has been verified to be useful in supervised VI-ReID solutions [54]. At each training epoch, DBSCAN [12] is used to cluster images within each modality. Following [45], the maximum distance for DBSCAN is set to 0.6 on SYSU and 0.3 on RegDB. The minimal number for two datasets is set to 4 during clustering. Following [45], the memory updating rate λ is 0.1 and the temperature factor τ is 0.05. The weighting parameter μ is 0.5. This work is supported by Huawei MindSpore [20].

4.2. Comparison with the State-of-the-Arts

To demonstrate the efficiency of our proposed methods, we compare it with three related Re-ID settings, which are supervised VI-ReID (SVI-ReID), unsupervised learn-

	SYSU-MM01 Settings		All search					Indoor Search				
	Methods	Venue	$r1(\%)$	$r10(\%)$	$r20(\%)$	mAP(%)	mINP(%)	$r1(\%)$	$r10(\%)$	$r20(\%)$	mAP(%)	mINP(%)
SVI-ReID	Hi-CMD [8]	CVPR'20	34.9	77.60	-	35.9	-	-	-	-	-	-
	DDAG [52]	ECCV'20	54.75	90.39	95.81	53.02	39.62	61.02	94.06	98.41	67.98	62.21
	AGW [53]	TPAMI'21	47.5	84.39	92.14	47.65	35.3	54.17	91.94	95.98	62.97	59.23
	LbA [29]	ICCV'21	55.41	-	-	54.14	-	58.46	-	-	66.33	-
	CAJ [51]	ICCV'21	69.88	95.71	98.46	66.89	53.61	76.26	97.88	99.49	80.37	76.79
	MPANet [40]	CVPR'21	70.58	96.21	98.80	68.24	-	76.74	98.21	99.57	80.95	-
	FMCNet [55]	CVPR'22	66.34	-	-	62.51	-	68.15	-	-	74.09	-
	DART [47]	CVPR'22	68.72	96.36	98.96	66.29	53.26	72.52	97.84	99.46	78.17	74.94
USL-ReID	SPCL [15]	NIPS'20	18.37	54.08	69.02	19.39	10.99	26.83	68.31	83.24	36.42	33.05
	MMT [14]	ICLR'20	21.47	59.65	73.29	21.53	11.50	22.79	63.18	79.04	31.50	27.66
	ICE [4]	ICCV'21	20.54	57.50	70.89	20.39	10.24	29.81	69.41	82.66	38.35	34.32
	IICS‡ [42]	CVPR'21	14.39	47.91	62.32	15.74	8.41	15.91	54.20	71.49	24.87	22.15
	PPLR‡ [7]	CVPR'22	11.98	43.17	59.02	12.25	4.97	12.71	48.66	68.76	20.81	17.61
USVI-ReID	H2H [22]	TIP'21	30.15	65.92	77.32	29.40	-	-	-	-	-	-
	OTLA [33]	ECCV'22	29.9	-	-	27.1	-	29.8	-	-	38.8	-
	OTLA(SS†)	ECCV'22	48.2	-	-	43.9	-	47.4	-	-	56.8	-
	ADCA [45]	MM'22	45.51	85.29	93.16	42.73	28.29	50.60	89.66	96.15	59.11	55.17
	ADCA(AGW)	MM'22	50.90	88.98	95.97	45.70	29.12	51.39	90.14	95.29	59.82	56.08
	Ours	-	**57.27**	**92.48**	**97.23**	**51.78**	**34.96**	**56.23**	**90.19**	**95.39**	**62.74**	**58.13**

Table 1. Comparison with the state-of-the-art methods on SYSU-MM01. † indicates semi-supervised setting when the method utilizes the visible label. ‡ indicates we re-implement the result with the official code. Rank at r accuracy(%), mAP (%) and mINP (%) are reported.

	RegDB	Visible to Thermal		Thermal to Visible	
	Methods	$r1(\%)$	mAP(%)	$r1(\%)$	mAP(%)
SVI-ReID	Hi-CMD [8]	70.93	66.04	-	-
	DDAG [52]	69.34	63.46	68.06	61.80
	AGW [53]	70.05	66.37	70.49	65.90
	LbA [29]	74.17	67.64	72.43	65.46
	CAJ [51]	85.03	79.14	84.75	77.82
	MPANet [40]	83.70	80.90	82.80	80.70
	FMCNet [55]	89.12	84.43	88.38	83.86
	DART [47]	83.60	75.67	81.97	73.38
USL-ReID	SPCL [15]	13.59	14.68	11.70	13.56
	MMT [14]	25.68	26.51	24.42	25.59
	ICE [4]	12.98	15.64	12.18	14.82
	IICS [42]	9.17	9.94	9.11	9.90
	PPLR [7]	10.30	11.94	10.39	11.23
USVI-ReID	H2H [22]	23.81	18.87	-	-
	OTLA [33]	32.90	29.70	32.10	28.60
	OTLA(SS)	49.90	41.80	49.60	42.80
	ADCA [45]	67.20	64.05	68.48	63.81
	ADCA(AGW)	66.62	63.47	67.29	62.98
	Ours	**69.48**	**65.41**	**69.85**	**65.17**

Table 2. Comparison with the state-of-the-art methods on RegDB. Rank at r accuracy(%), mAP (%) and mINP (%) are reported.

ing ReID (USL-ReID) and unsupervised VI-ReID (USVI-ReID), respectively. The results on SYSU-MM01 and RegDB are shown in Tab. 1 and Tab. 2.

Comparison with SVI-ReID Methods. It is encouraging that our proposed unsupervised approach can outperform some recent supervised SVI-ReID methods (see DDAG [52] and AGW [53] on SYSU-MM01). This phenomenon indicates that reliable cross-modality correspondences can be obtained by progressive graph matching. It must be acknowledged that there is still much room for improvement for unsupervised methods compared to their supervised counterparts due to the absence of annotated cross-modality correspondences.

Comparison with USL-ReID Methods. The results shown in Tab. 1 and Tab. 2 indicate USL-ReID methods can not bridge the large modality gap in cross-modality ReID. It is unfair to directly compare these methods with our solution. We list them here to demonstrate the necessity of proposing specific solutions for cross-modality scenarios.

Comparison with USVI-ReID Methods. There are few USVI-ReID methods as this is a relatively new task. We select all three reported methods for comparison. OTLA and H2H both need an extra annotated visible dataset. Compared with them, ADCA is more similar to our approach, since we both try to handle pure unsupervised VI-ReID task. In Tab. 1 and Tab. 2, we can see that our method is significantly better than all existing USVI-ReID methods, demonstrating the effectiveness of our method.

4.3. Ablation Study

We conduct ablation study in this subsection to validate the effectiveness of each component of our method. Firstly, we explain the different settings in this study. DCL is the baseline described in Sec. 3.1. BGM and PGM indicate which matching method is selected and are both defined in Sec. 3.2. DM (Direct Matching) refers to the method proposed in [45], which is designed to find positive cross-modality pairs, also. $\mathcal{L}_{ccl}$ and $\mathcal{L}_{accl}$ indicate which loss function is adopted. The main results are shown in Tab. 3.

Effectiveness of PGM. To verify the effectiveness of the PGM method, we first calculate the accuracy of cross-modality correspondences, and the results of different matching methods are shown in Fig. 4. The matching ac-

Order	Components							SYSU-MM01 (All)			SYSU-MM01 (Indoor)			RegDB (Visible to Infrared)		
	DCL	BGM†	PGM	$\mathcal{L}_{R2V}$	$\mathcal{L}_{V2R}$	$\mathcal{L}_{ccl}$	$\mathcal{L}_{accl}$	$r1$	mAP	mINP	$r1$	mAP	mINP	$r1$	mAP	mINP
1	✓							39.98	39.36	26.43	45.95	53.83	49.62	43.78	42.50	32.21
2	✓	✓				✓		48.82	43.32	26.83	50.51	58.41	54.32	67.21	62.87	49.35
3	✓	✓					✓	51.98	47.39	31.08	53.42	61.04	56.73	67.50	63.91	51.19
4	✓		✓	✓				45.52	41.58	25.65	45.34	54.34	49.97	66.34	61.78	18.98
5	✓		✓		✓			51.12	45.88	29.03	54.13	59.08	55.40	67.38	63.12	50.08
6	✓		✓			✓		52.25	47.74	31.88	53.16	60.88	56.72	67.42	63.50	51.20
7	✓		✓				✓	**57.27**	**51.78**	**34.96**	**56.23**	**62.74**	**58.13**	**69.48**	**65.41**	**52.97**

Table 3. Ablation studies on SYSU-MM01 and RegDB. † indicates we utilize basic graph matching and instances without correspondences are not involved in cross-modality contrastive learning. Rank at r accuracy(%), mAP (%) and mINP (%) are reported.

curacy is defined as the ratio of the cluster whose own label matches its corresponding label to all the clusters of the modality. Besides, the label of the cluster is determined by the label of the largest number of instances in the cluster. We find that though a bit low in the first epochs, the matching accuracy can be gradually improved by the proposed PGM method (see Fig. 4). The comparison between PGM (BGM) and DM validates that the former can produce more reliable correspondences with global consideration. Furthermore, by comparing PGM with BGM (see 2^{nd} row and 4^{th} row, 3^{rd} row and 5^{th} row in Tab. 3), we believe that the progressive strategy handles better the cluster imbalance issue.

Effectiveness of ACCL. We must state that the ACCL module does not function alone because it relies on the previously generated cross-modality correspondences. When combined ACCL with PGM, it boosts the performance of unsupervised setting (see 1^{st} row and 5^{th} row in Tab. 3). This implies that ACCL makes efficient use of cross-modality correspondences and decreases modal differences effectively. When comparing ACCL with CCL (see 2^{nd} row and 3^{rd} row, 4^{th} row and 5^{th} row), the former demonstrates its advantages. This is because the alternate scheme is more reasonable in dealing with noisy correspondences.

4.4. Parameters Analysis

The proposed method includes the parameter μ, which is the weighting parameter to combine different losses. In this experiment, we aim to study how much the ACCL would affect the performance. First, we test our model with varying μ over a range $\{0, 0.3, 0.5, 0.7, 1.0\}$ on SYSU-MM01. When $\mu = 0$ (ACCL is not used), the model suffers poor performance. We also observe that the performance is not sensitive to μ since the results are high and relatively similar in all the cases. Comparing the results with different μ, we can conclude that our proposed ACCL can effectively improve the performance of the model.

5. Conclusion

In this paper, we propose the progressive graph matching method to find reliable cross-modality correspondences for unsupervised VI-ReID. We first formulate the correspondences mining procedure as graph matching process with global consideration and progressively utilize graph matching to handle the cluster imbalance issue. Moreover, an alternate cross contrastive learning is designed to learn modality-invariant features. Extensive experiments demonstrate the state-of-the-art performance of our method.

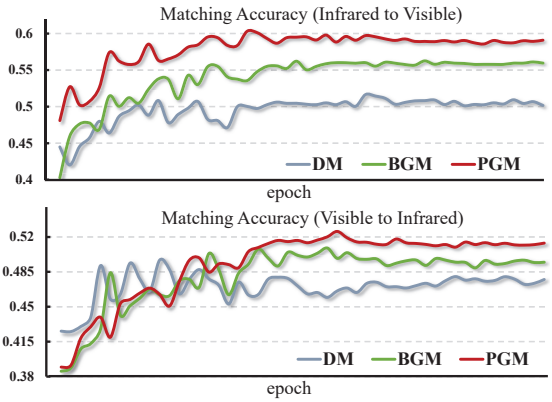

Figure 4. Matching accuracy on SYSU-MM01 dataser under the *all-search* mode.

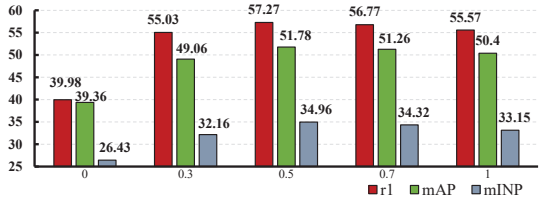

Figure 5. The effect of parameter μ on SYSU-MM01 dataset under the *all-search* mode. μ is used to balance $\mathcal{L}_{dcl}$ and $\mathcal{L}_{accl}$.

Acknowledgement. This work is partially supported by the Key Research and Development Program of Hubei Province (2021BAA187), National Natural Science Foundation of China under Grant (62176188), Zhejiang lab (NO.2022NF0AB01), the Special Fund of Hubei Luojia Laboratory (220100015) and CAAI-Huawei MindSpore Open Fund.

References

[1] Zechen Bai, Zhigang Wang, Jian Wang, Di Hu, and Errui Ding. Unsupervised multi-source domain adaptation for person re-identification. In *CVPR*, pages 12914–12923, 2021. 3

[2] Derek Bruff. The assignment problem and the hungarian method. *Notes for Math*, 20(29-47):5, 2005. 5

[3] 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*, 31:2352–2364, 2022. 1

[4] Hao Chen, Benoit Lagadec, and Francois Bremond. Ice: Inter-instance contrastive encoding for unsupervised person re-identification. In *ICCV*, pages 14960–14969, 2021. 2, 7

[5] Yehansen Chen, Lin Wan, Zhihang Li, Qianyan Jing, and Zongyuan Sun. Neural feature search for rgb-infrared person re-identification. In *CVPR*, pages 587–597, 2021. 1

[6] Minsu Cho and Kyoung Mu Lee. Progressive graph matching: Making a move of graphs via probabilistic voting. In *CVPR*, pages 398–405, 2012. 2

[7] 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. 2, 3, 7

[8] 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*, June 2020. 7

[9] Zuozhuo Dai, Guangyuan Wang, Weihao Yuan, Siyu Zhu, and Ping Tan. Cluster contrast for unsupervised person re-identification. In *ACCV*, pages 1142–1160, 2022. 2, 3, 4

[10] 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. 6

[11] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *CVPR*, pages 994–1003, 2018. 3

[12] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, pages 226–231, 1996. 3, 6

[13] 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. 3

[14] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. In *ICLR*, 2019. 3, 7

[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*, 33:11309–11321, 2020. 3, 7

[16] Seyed Hamid Rezatofighi, Anton Milan, Zhen Zhang, Qinfeng Shi, Anthony Dick, and Ian Reid. Joint probabilistic matching using m-best solutions. In *CVPR*, pages 136–145, 2016. 3

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 3, 6

[18] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 3

[19] Zhipeng Huang, Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Peng Chu, Quanzeng You, Jiang Wang, Zicheng Liu, and Zheng-jun Zha. Lifelong unsupervised domain adaptive person re-identification with coordinated anti-forgetting and adaptation. In *CVPR*, pages 14288–14297, 2022. 3

[20] Huawei. Mindspore, https://www.mindspore.cn/, 2020. 6

[21] Kongzhu Jiang, Tianzhu Zhang, Xiang Liu, Bingqiao Qian, Yongdong Zhang, and Feng Wu. Cross-modality transformer for visible-infrared person re-identification. In *ECCV*, pages 480–496, 2022. 3

[22] Wenqi Liang, Guangcong Wang, Jianhuang Lai, and Xiaohua Xie. Homogeneous-to-heterogeneous: Unsupervised learning for rgb-infrared person re-identification. *IEEE TIP*, 30:6392–6407, 2021. 2, 3, 7

[23] Xinyu Lin, Jinxing Li, Zeyu Ma, Huafeng Li, Shuang Li, Kaixiong Xu, Guangming Lu, and David Zhang. Learning modal-invariant and temporal-memory for video-based visible-infrared person re-identification. In *CVPR*, pages 20973–20982, 2022. 1, 2, 3

[24] Yutian Lin, Xuanyi Dong, Liang Zheng, Yan Yan, and Yi Yang. A bottom-up clustering approach to unsupervised person re-identification. In *AAAI*, pages 8738–8745, 2019. 3

[25] 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. 1, 2, 3, 6

[26] Djebril Mekhazni, Amran Bhuiyan, George Ekladious, and Eric Granger. Unsupervised domain adaptation in the dissimilarity space for person re-identification. In *ECCV*, pages 159–174, 2020. 3

[27] 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. 6

[28] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE TKDE*, 22(10):1345–1359, 2009. 3

[29] Hyunjong Park, Sanghoon Lee, Junghyup Lee, and Bumsub Ham. Learning by aligning: Visible-infrared person re-identification using cross-modal correspondences. In *ICCV*, pages 12046–12055, October 2021. 1, 2, 7

[30] Seyed Hamid Rezatofighi, Anton Milan, Zhen Zhang, Qinfeng Shi, Anthony Dick, and Ian Reid. Joint probabilistic data association revisited. In *ICCV*, pages 3047–3055, 2015. 5

[31] Dongkai Wang and Shiliang Zhang. Unsupervised person re-identification via multi-label classification. In *CVPR*, pages 10981–10990, 2020. 3

[32] Guan'an Wang, Tianzhu Zhang, Jian Cheng, Si Liu, Yang Yang, and Zengguang Hou. Rgb-infrared cross-modality per-

son re-identification via joint pixel and feature alignment. In *ICCV*, pages 3623–3632, 2019. 3

[33] Jiangming Wang, Zhizhong Zhang, Mingang Chen, Yi Zhang, Cong Wang, Bin Sheng, Yanyun Qu, and Yuan Xie. Optimal transport for label-efficient visible-infrared person re-identification. In *ECCV*, pages 93–109, 2022. 2, 3, 7

[34] 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. 2

[35] Siwei Wang, Xinwang Liu, Suyuan Liu, Jiaqi Jin, Wenxuan Tu, Xinzhong Zhu, and En Zhu. Align then fusion: Generalized large-scale multi-view clustering with anchor matching correspondences. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *NeurIPS*, 2022. 2

[36] 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. 3

[37] 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. 3

[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. 1, 2, 6

[39] Jinlin Wu, Yang Yang, Hao Liu, Shengcai Liao, Zhen Lei, and Stan Z Li. Unsupervised graph association for person re-identification. In *ICCV*, pages 8321–8330, 2019. 3

[40] 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. 1, 2, 3, 7

[41] Yuanlu Xu, Liang Lin, Wei-Shi Zheng, and Xiaobai Liu. Human re-identification by matching compositional template with cluster sampling. In *ICCV*, pages 3152–3159, 2013. 3

[42] Shiyu Xuan and Shiliang Zhang. Intra-inter camera similarity for unsupervised person re-identification. In *CVPR*, pages 11926–11935, 2021. 2, 7

[43] Shiyu Xuan and Shiliang Zhang. Intra-inter domain similarity for unsupervised person re-identification. *IEEE TPAMI*, pages 1–1, 2022. 3

[44] Junchi Yan, Minsu Cho, Hongyuan Zha, Xiaokang Yang, and Stephen M Chu. Multi-graph matching via affinity optimization with graduated consistency regularization. *IEEE TPAMI*, 38(6):1228–1242, 2015. 2

[45] Bin Yang, Mang Ye, Jun Chen, and Zesen Wu. Augmented dual-contrastive aggregation learning for unsupervised visible-infrared person re-identification. In *ACM MM*, pages 2843–2851, 2022. 2, 3, 6, 7

[46] Fengxiang Yang, Zhun Zhong, Zhiming Luo, Yuanzheng Cai, Yaojin Lin, Shaozi Li, and Nicu Sebe. Joint noise-tolerant learning and meta camera shift adaptation for unsupervised person re-identification. In *CVPR*, pages 4855–4864, 2021. 3

[47] Mouxing Yang, Zhenyu Huang, Peng Hu, Taihao Li, Jiancheng Lv, and Xi Peng. Learning with twin noisy labels for visible-infrared person re-identification. In *CVPR*, pages 14308–14317, 2022. 2, 3, 7

[48] Mang Ye, Xiangyuan Lan, Qingming Leng, and Jianbing Shen. Cross-modality person re-identification via modality-aware collaborative ensemble learning. *IEEE TIP*, 29:9387–9399, 2020. 6

[49] Mang Ye, Jiawei Li, Andy J. Ma, Liang Zheng, and Pong C. Yuen. Dynamic graph co-matching for unsupervised video-based person re-identification. *IEEE Transactions on Image Processing*, 28(6):2976–2990, 2019. 2

[50] Mang Ye, Andy J Ma, Liang Zheng, Jiawei Li, and Pong C Yuen. Dynamic label graph matching for unsupervised video re-identification. In *ICCV*, pages 5142–5150, 2017. 2, 3

[51] Mang Ye, Weijian Ruan, Bo Du, and Mike Zheng Shou. Channel augmented joint learning for visible-infrared recognition. In *ICCV*, pages 13567–13576, 2021. 1, 2, 3, 6, 7

[52] 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. 1, 2, 6, 7

[53] 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*, 44(6):2872–2893, 2021. 3, 6, 7

[54] Mang Ye, Jianbing Shen, and Ling Shao. Visible-infrared person re-identification via homogeneous augmented tri-modal learning. *IEEE Transactions on Information Forensics and Security*, 16:728–739, 2020. 6

[55] 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. 3, 7

[56] Xiao Zhang, Yixiao Ge, Yu Qiao, and Hongsheng Li. Refining pseudo labels with clustering consensus over generations for unsupervised object re-identification. In *CVPR*, pages 3436–3445, 2021. 3

[57] Xinyu Zhang, Dongdong Li, Zhigang Wang, Jian Wang, Errui Ding, Javen Qinfeng Shi, Zhaoxiang Zhang, and Jingdong Wang. Implicit sample extension for unsupervised person re-identification. In *CVPR*, pages 7369–7378, 2022. 2, 3

[58] Yiyuan Zhang, Sanyuan Zhao, Yuhao Kang, and Jianbing Shen. Modality synergy complement learning with cascaded aggregation for visible-infrared person re-identification. In *ECCV*, pages 462–479, 2022. 3

[59] Ziming Zhang and Venkatesh Saligrama. Prism: Person reidentification via structured matching. *IEEE TCSVT*, 27(3):499–512, 2016. 3

[60] Kecheng Zheng, Wu Liu, Lingxiao He, Tao Mei, Jiebo Luo, and Zheng-Jun Zha. Group-aware label transfer for domain adaptive person re-identification. In *CVPR*, pages 5310–5319, 2021. 3

[61] Yi Zheng, Shixiang Tang, Guolong Teng, Yixiao Ge, Kaijian Liu, Jing Qin, Donglian Qi, and Dapeng Chen. Online pseudo label generation by hierarchical cluster dynamics

for adaptive person re-identification. In *ICCV*, pages 8371–8381, 2021. 3

[62] Zhun Zhong, Liang Zheng, Shaozi Li, and Yi Yang. Generalizing a person retrieval model hetero-and homogeneously. In *ECCV*, pages 172–188, 2018. 3

[63] 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. 3

[64] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Learning to adapt invariance in memory for person re-identification. *IEEE TPAMI*, 43(8):2723–2738, 2020. 3