Learning Commonality, Divergence and Variety for Unsupervised Visible-Infrared Person Re-identification

Jiangming Shi[1,3]*, **Xiangbo Yin**[2*], **Yachao Zhang**[2], **Zhizhong Zhang**[4,5]
Yuan Xie[3,4†], **Yanyun Qu**[1,2†]
[1]Institute of Artificial Intelligence, Xiamen University
[2]School of Informatics, Xiamen University
[3]Shanghai Innovation Institute
[4]East China Normal University
[5]Shanghai Key Laboratory of Computer Software Evaluating and Testing
jiangming.shi@outlook.com yxie@cs.ecnu.edu.cn yyqu@xmu.edu.cn
Code: https://github.com/shijiangming1/PCLHD

Abstract

Unsupervised visible-infrared person re-identification (USVI-ReID) aims to match specified persons in infrared images to visible images without annotations, and vice versa. USVI-ReID is a challenging yet underexplored task. Most existing methods address the USVI-ReID through cluster-based contrastive learning, which simply employs the cluster center to represent an individual. However, the cluster center primarily focuses on commonality, overlooking divergence and variety. To address the problem, we propose a Progressive Contrastive Learning with Hard and Dynamic Prototypes for USVI-ReID. In brief, we generate the hard prototype by selecting the sample with the maximum distance from the cluster center. We reveal that the inclusion of the hard prototype in contrastive loss helps to emphasize divergence. Additionally, instead of rigidly aligning query images to a specific prototype, we generate the dynamic prototype by randomly picking samples within a cluster. The dynamic prototype is used to encourage variety. Finally, we introduce a progressive learning strategy to gradually shift the model's attention towards divergence and variety, avoiding cluster deterioration. Extensive experiments conducted on the publicly available SYSU-MM01 and RegDB datasets validate the effectiveness of the proposed method.

1 Introduction

Visible-infrared person re-identification (VI-ReID) aims at matching the same person captured in one modality with their counterparts in another modality [1–3]. It has recently gained attention in computer vision applications like video surveillance [4] and image retrieval [5–7]. With the development of deep learning [8–11], VI-ReID has achieved remarkable advancements [12–14]. However, the development of existing VI-ReID methods is still limited due to the requirement for expensive-annotated training data [15, 16]. To mitigate the problem of annotating large-scale cross-modality data, some semi-supervised VI-ReID methods [17–19] are proposed to learn modality-invariant and identity-related discriminative representations by utilizing both labeled and unlabeled data. For this purpose, OTLA [17] proposed an optimal transport label assignment mechanism to assign pseudo-labels for unlabeled infrared images while ignoring how to calibrate noise pseudo-labels. DPIS [18] integrates two pseudo-labels generated by distinct models into a hybrid pseudo-label

*Equal contribution.
†Corresponding author.

38th Conference on Neural Information Processing Systems (NeurIPS 2024), Vancouver.

 https://doi.org/10.52202/079017-3163

for unlabeled infrared data, but it makes the training process more complex. Although these methods have gained promising performances, they still rely on a certain number of manual-labeled data.

Several USVI-ReID methods [20–23] have proposed to tackle the issues of expensive visible-infrared annotation through contrastive learning. These methods create two modality-specific memories, one for visible features and the other for infrared features. During training, these methods consider the memory center as a prototype and minimize the contrastive loss across the features of query images and prototype. Then, these methods aggregate the corresponding prototypes based on similarity. However, the centroid prototype only stores the commonality of each person, neglecting the divergence [24–26], which causes the pseudo-labels generated by the cluster to be unreliable. Just like a normal distribution, to better reflect the data distribution of a dataset, we need not only the mean but also the variance.

In this paper, we argue that an important aspect of contrastive learning for USVI-ReID, i.e. the design of the prototype, has so far been neglected, and propose progressive contrastive learning with hard and dynamic prototype (PCLHD) method for the USVI-ReID. Firstly, we design a Hard Prototype Contrastive Learning (HPCL) to mine divergent yet meaningful information. In contrast to traditional contrastive learning methods, we choose the hard samples to serve as the hard prototype. In other words, the hard prototype is the one that is farthest from the memory center. The hard prototype encompasses distinctive information. Furthermore, we introduce the concept of Dynamic Prototype Contrastive Learning (DPCL), we randomly select samples from each cluster to serve as the dynamic prototype. DPCL effectively accounts for the intrinsic variety within clusters, enhancing the model's adaptability to varying data distributions. Early clustering results are unreliable, and utilizing hard and dynamic prototype at this stage may lead to cluster degradation. Therefore, we introduce progressive contrastive learning to gradually focus on divergence and variety.

The main contributions are summarized as follows:

- We propose a progressive contrastive learning with hard and dynamic prototype method for the USVI-ReID. We reconsider the design of prototypes in contrastive learning to ensure that the model stably captures commonality, divergence, and variety.
- We propose Hard Prototype Contrastive Learning for mining divergent yet significant information, and Dynamic Prototype Contrastive Learning for preserving the intrinsic variety in sample features.
- Experiments on SYSU-MM01 and RegDB datasets demonstrate the superiority of our method compared to existing USVI-ReID methods, and PCLHD generates higher-quality pseudo-labels than other methods.

2 Related Work

2.1 Supervised Visible-Infrared Person ReID

Visible-infrared person re-identification (VI-ReID) has drawn much attention in recent years [27–32]. Many VI-ReID methods focused on mitigating huge semantic gaps across modalities have made advanced progress, which can be classified into two primary classes based on their different aligning ways: image-level alignment and feature-level alignment. The image-level alignment methods focus on reducing cross-modality gaps by modality translation. Some GAN-based methods [33, 34] are proposed to perform style transformation for aligning cross-modality images. However, the generated images unavoidably contain noise. Therefore, X-modality [35] and its promotions [36, 37] align cross-modality images by introducing a middle modality. Mainstream feature-level alignment methods [38–40] focus on minimizing cross-modality gaps by finding a modality-shared feature space. However, the advanced performances of the above methods build on large-scale human-labeled cross-modality data, which are quite time-consuming and expensive, thus hindering the fast application of these methods in real-scenes.

2.2 Unsupervised Single-Modality Person ReID

The existing unsupervised single-modality person ReID methods can be roughly divided into two classes: Unsupervised domain adaption (UDA) methods, which try to leverage the knowledge transferred from labeled source domain to improve performance [41–44], and fully unsupervised

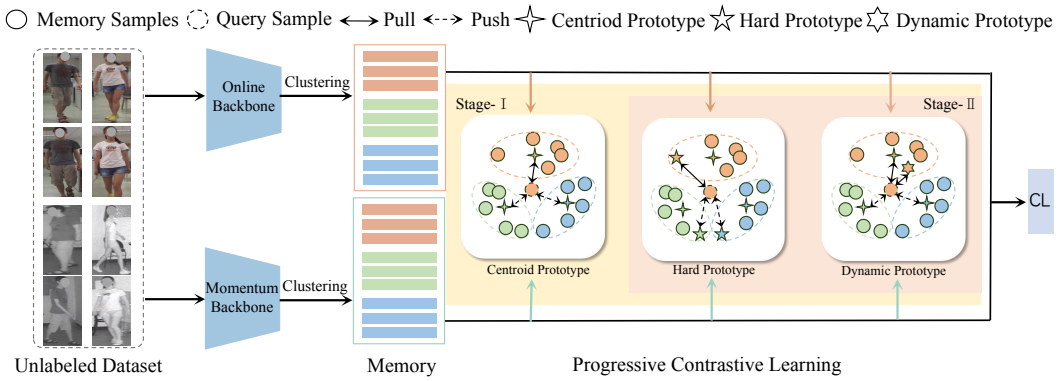

Figure 1: Framework of our PCLHD. The framework consists of two stages: the first stage employs contrastive learning with centroid prototypes to learn well-discriminative representation, and the second stage introduces contrastive learning with hard and dynamic prototypes to further focus on divergence and variety.

methods (USL), which directly train a USL-ReID model on the unlabeled target domain [21, 45]. Compared with the UDA methods, the USL methods are more challenging. Recently, cluster-contrast learning [46] has achieved impressive performance by performing contrastive learning at the cluster level. However, cluster-contrast with a uni-proxy can be biased and confusing, which fails to accurately describe the information of a cluster. To this end, the methods [47, 48] proposed maintaining multi-proxies for a cluster to adaptively capture different information within the cluster. The above methods are mainly proposed to solve the single-modality ReID task, but they are limited to solving the USL-VI-ReID task due to large cross-modality gaps.

2.3 Unsupervised Visible-Infrared Person ReID

Unsupervised visible-infrared person ReID (USVI-ReID) has attracted much attention due to the advantage of not relying on any data annotation. Some UDA methods [49, 17] use a well-annotated labeled source domain for pre-training to solve the USVI-ReID task. Some fully unsupervised methods [23, 22] adopt contrastive learning to boost performance, which mainly follow a two-step loop paradigm: generating pseudo-labels using the DBSCAN algorithm [50] to create memory banks with clustering centers and establishing cross-modality correspondences based on these memory banks. However, pseudo-labels are often inaccurate and rigid, CCLNet [51] leverages the text information from CLIP to afford greater semantic monitoring insights to compensate for the rigidity of pseudo-labels. Moreover, reliable cross-modality correspondences are vital to USVI-ReID, thus PGM [23] proposes a progressive graph matching framework to establish more reliable cross-modality correspondences. However, cluster centers mainly present common information while lacking distinctive information, which results in ambiguous cross-modality correspondences when meeting hard samples [52, 53].

3 Method

3.1 Problem Formulation and Overview

Given a USVI-ReID dataset $D = \{V, R\}$, where $V = \{V_i\}_{i=1}^{N_v}$ represents the visible images and $R = \{R_j\}_{j=1}^{N_r}$ denotes the infrared images. V_i and R_j represent the set of images corresponding to the i-th and j-th class. N_v and N_r denote the number of visible and infrared clusters, respectively. In the USVI-ReID task, the purpose is to train a deep neural network to obtain modality-invariant and identity-related features for matching pedestrian images with the same identity.

We propose a Progressive Contrastive Learning with Hard and Dynamic Prototype (PCLHD) method for USVI-ReID, which mainly contains online encoder, momentum encoder, and progressive contrastive learning strategy with centroid prototype, hard prototype, and dynamic prototype, as shown in Fig. 1. The online encoder is a standard network, updated through back-propagation. The momentum

https://doi.org/10.52202/079017-3163

encoder mirrors the structure of the online encoder, updated through the weights of the online encoder. The clustering is used to generate pseudo labels for creating cluster-aware memory, and we employ DBSCAN for clustering. PCLHD primarily focuses on representation learning, and we use PGM [23] to aggregate cross-modality memory.

3.2 Centroid Prototype Contrastive Learning

Following the USVI-ReID methods [22, 54], we use centroid prototype contrastive learning to optimize the online encoder in the first state, which includes memory initialization and optimization.

Memory Initialization. Let ϕ_0 be the online encoder that transforms the input image to an embedding vector. At the beginning of each training epoch, all image features are clustered by DBSCAN [50] and then each cluster's representations are stored in visible memory M_{RGB} =$\{cm_1^v, cm_2^v, \cdots, cm_{N_v}^v\}$ and infrared memory M_{IR}=$\{cm_1^r, cm_2^r, \cdots, cm_{N_r}^r\}$, as follows:

$$cm_i^v = \frac{1}{|V_i|} \sum_{\mathrm{v} \in V_i} \phi_0(\mathrm{v}), \tag{1}$$

$$cm_j^r = \frac{1}{|R_j|} \sum_{\mathrm{r} \in R_j} \phi_0(\mathrm{r}), \tag{2}$$

where $|\cdot|$ denotes the number of instances belonging to specific cluster.

Optimization. During training, we update the two modality-specific memories by a momentum updating strategy [46]. We treat the memory center as a centroid prototype and optimize the feature extractor ϕ_0 using contrastive learning with the centroid prototype, computed as:

$$\mathcal{L}_{CPCL}^v = \frac{1}{N_v} \sum_{i \in N_v} \frac{-1}{|V_i|} \sum_{\mathrm{v} \in V_i} \log \frac{\exp\left(\phi_0(\mathrm{v}) \cdot cm_i^v / \tau\right)}{\sum_{j \in N_v} \exp\left(\phi_0(\mathrm{v}) \cdot cm_j^v / \tau\right)}, \tag{3}$$

$$\mathcal{L}_{CPCL}^r = \frac{1}{N_r} \sum_{i \in N_r} \frac{-1}{|R_i|} \sum_{\mathrm{r} \in R_i} \log \frac{\exp\left(\phi_0(\mathrm{r}) \cdot cm_i^r / \tau\right)}{\sum_{j \in N_r} \exp\left(\phi_0(\mathrm{r}) \cdot cm_j^r / \tau\right)}, \tag{4}$$

$$\mathcal{L}_{CPCL} = \mathcal{L}_{CPCL}^v + \mathcal{L}_{CPCL}^r, \tag{5}$$

where $cm_i^{v(r)}$ is the positive centroid prototype, denoting a query and the prototype shares the same identity. The τ is a temperature hyper-parameter.

3.3 Hard Prototype Contrastive Learning

To ensure that the prototype effectively captures divergence within a identity, we devise a novel hard prototype for contrastive learning, which is referred to as Hard Prototype Contrastive Learning. HPCL is designed to provide a comprehensive understanding of personal characteristics, which benefits its handling of hard samples [47]. We use the online encoder ϕ_0 to extract feature representations, and select k samples that are farthest from the memory center as the hard prototype:

$$hm_i^v = \underset{\forall \mathrm{v} \in V_i}{\arg\max} \left\| \phi_0(\mathrm{v}) - cm_i^v \right\|, \tag{6}$$

$$hm_j^r = \underset{\forall \mathrm{r} \in R_j}{\arg\max} \left\| \phi_0(\mathrm{r}) - cm_j^r \right\|. \tag{7}$$

Theorem 1. The information entropy of hard sample prototypes is greater than the information entropy of centroid prototypes, thereby preserving greater divergence within the hard memory.

Given a set of features $\{f_1, f_2, \ldots, f_{N_c}\}$ for class c. The entropy $H(cm_c)$ can be approximated by the entropy of the distribution of the sample means. Considering that cm_c is a convex combination of the sample features f_i, we have:

$$H\left(cm_c\right) = H\left(\frac{1}{N_c}\sum_{i=1}^{N_c} f_i\right). \tag{8}$$

By the convexity of entropy, we have:

$$H\left(\frac{1}{N_c}\sum_{i=1}^{N_c} f_i\right) \leq \frac{1}{N_c}\sum_{i=1}^{N_c} H(f_i). \tag{9}$$

This inequality implies that the entropy of the centroid prototype is generally lower due to the averaging effect, which reduces the divergence among the samples, leading to lower entropy. Given that hm_c is the sample with the maximum individual entropy among the set $\{f_1, f_2, \ldots, f_{N_c}\}$, it follows that:

$$H(hm_c) \geq \frac{1}{N_c}\sum_{i=1}^{N_c} H(f_i) \geq H(cm_c). \tag{10}$$

Then, we construct contrastive loss with the hard prototype to minimize the distance between the query and the positive hard prototype while maximizing their discrepancy to all other cluster hard prototypes, as follows:

$$\mathcal{L}^v_{HPCL} = \frac{1}{N_v}\sum_{i\in N_v}\frac{-1}{|V_i|}\sum_{\mathrm{v}\in V_i}\log\frac{\exp\left(\phi_0(\mathrm{v})\cdot hm_i^v/\tau\right)}{\sum\limits_{j\in N_v}\exp\left(\phi_0(\mathrm{v})\cdot hm_j^v/\tau\right)}, \tag{11}$$

$$\mathcal{L}^r_{HPCL} = \frac{1}{N_r}\sum_{i\in N_r}\frac{-1}{|R_i|}\sum_{\mathrm{r}\in R_i}\log\frac{\exp\left(\phi_0(\mathrm{r})\cdot hm_i^r/\tau\right)}{\sum\limits_{j\in N_r}\exp\left(\phi_0(\mathrm{r})\cdot hm_j^r/\tau\right)}, \tag{12}$$

$$\mathcal{L}_{HPCL} = \mathcal{L}^v_{HPCL} + \mathcal{L}^r_{HPCL}, \tag{13}$$

where $hm_i^{v(r)}$ is the positive hard prototype representation and the τ is a temperature hyper-parameter.

Finally, we update the two modality-specific memories with a momentum-updating strategy:

$$hm^v_{i,t} = \alpha hm^v_{i,t-1} + (1-\alpha)\phi_0(\mathrm{v}), \forall \mathrm{v} \in V_i \tag{14}$$

$$hm^r_{i,t} = \alpha hm^r_{i,t-1} + (1-\alpha)\phi_0(\mathrm{r}), \forall \mathrm{r} \in R_i \tag{15}$$

where α is a momentum coefficient that controls the update speed of the memories. t and $t-1$ refer to the current and last iteration, respectively.

The hard prototype contrastive learning has two main advantages: For intra-class feature learning, it ensures that the learning process does not just focus on the shared characteristics within a cluster but also considers the diverse elements, which are often more informative. For inter-class feature learning, it is also beneficial for increasing the distances between different persons. In contrast, centroid prototypes tend to average features, lacking diversity, which can affect the network's ability to extract discriminative features.

3.4 Dynamic Prototype Contrastive Learning

Inspired by MoCo [55] and DPM [56], we design dynamic prototype contrastive learning in order to preserve the intrinsic variety in sample features. DPCL comprises an online encoder ϕ_0 and a momentum encoder ϕ_m. The momentum encoder mirrors the structure of the online encoder, which is updated by the accumulated weights of the online encoder:

$$\phi^t_m = \beta\phi^{t-1}_m + (1-\beta)\phi^t_0, \tag{16}$$

where β is a momentum coefficient that controls the update speed of the momentum encoder. t and $t-1$ refer to the current and last iteration, respectively. The momentum encoder ϕ_m is updated by the moving averaged weights, which are resistant to sudden fluctuations or noisy updates [55].

https://doi.org/10.52202/079017-3163

We use the momentum encoder ϕ_m to extract feature representation and store them in visible memory DM_{RGB}=$\{dm_1^v, dm_2^v, \cdots, dm_{N_v}^v\}$ and infrared memory DM_{IR}=$\{dm_1^r, dm_2^r, \cdots, dm_{N_r}^r\}$. We randomly select M visible/infrared samples from each cluster, denoted as X_i^v and X_j^r.as follows:

$$F_i^v = \phi_m(X_i^v), \tag{17}$$

$$F_j^r = \phi_m(X_j^r). \tag{18}$$

We select visible dynamic prototype dm_i^v from DM_{RGB}. In the same cluster, we select the sample farthest from the query image as the prototype. In different clusters, we choose the sample closest to the query image as the prototype:

$$dm_i^v = \begin{cases} \underset{\forall f_i^v \in F_i^v}{\arg\max} \|\phi_m(\mathrm{v}_j) - f_i^v\| & \text{if } y_j = y_i \\ \underset{\forall f_i^v \in F_i^v}{\arg\min} \|\phi_m(\mathrm{v}_j) - f_i^v\| & \text{if } y_j \neq y_i \end{cases}, \tag{19}$$

where y_q and y_i represent the pseudo label of the query image and the dynamic prototype, respectively. $\|\cdot\|$ denotes Euclidean norm. We obtain infrared prototype dm_j^r through the same method.

The overall optimization goal of DPCL is as follows:

$$\mathcal{L}_{DPCL}^v = \frac{1}{N_v} \sum_{i \in N_v} \frac{-1}{|V_i|} \sum_{\mathrm{v} \in V_i} \log \frac{\exp\left(\phi_m(\mathrm{v}) \cdot dm_i^v / \tau\right)}{\sum\limits_{j \in N_v} \exp\left(\phi_m(\mathrm{v}) \cdot dm_j^v / \tau\right)}, \tag{20}$$

$$\mathcal{L}_{DPCL}^r = \frac{1}{N_r} \sum_{i \in N_r} \frac{-1}{|R_i|} \sum_{\mathrm{r} \in R_i} \log \frac{\exp\left(\phi_m(\mathrm{r}) \cdot dm_i^r / \tau\right)}{\sum\limits_{j \in N_r} \exp\left(\phi_m(\mathrm{r}) \cdot dm_j^r / \tau\right)}, \tag{21}$$

$$\mathcal{L}_{DPCL} = \mathcal{L}_{DPCL}^v + \mathcal{L}_{DPCL}^r, \tag{22}$$

where $dm_i^{v(r)}$ is the positive dynamic prototype representation, i.e., the query image and dynamic prototype have the same identity.

DPCL promotes a flexible and adaptable learning process, aiming to minimize discrepancies between samples and their respective dynamic prototypes, rather than rigidly aligning query images with a fixed prototype.

3.5 Progressive Contrastive Learning

In the initial training phases, representations are generally of lower quality. Introducing hard samples at this period could be counterproductive, potentially leading the model optimization in an incorrect direction right from the start [47, 57]. To address this issue, we introduce the Progressive Contrastive Learning, which forms the overall loss function:

$$\mathcal{L}_{PCLHD} = \begin{cases} \mathcal{L}_{CPCL}, & \text{if epoch} \leqslant E_{\text{CPCL}} \\ \lambda \mathcal{L}_{HPCL} + (1-\lambda)\mathcal{L}_{DPCL}, & \text{else} \end{cases} \tag{23}$$

where λ is the loss weight, E_{CPCL} is a hyper-parameter.

4 Experiment

We conduct extensive experiments to validate the superiority of our proposed method. First, we provide the detailed experiment setting, which contains datasets, evaluation protocols, and implementation details. Then, we compare our method with many state-of-the-art VI-ReID methods and conduct ablation studies. In addition, to better illustrate our method, we also exhibit further analysis. If not specified, we conduct analysis experiments on SYSU-MM01 in the all-search mode.

Table 1: Comparisons with state-of-the-art methods on SYSU-MM01 and RegDB, including SVI-ReID, SSVI-ReID and USVI-ReID methods. All methods are measured by Rank-1 (%) and mAP (%). GUR* denotes the results without camera information.

Settings			SYSU-MM01				RegDB			
			All Search		Indoor Search		Visible2Thermal		Thermal2Visible	
Type	Method	Venue	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
SVI-ReID	DDAG [39]	ECCV'20	54.8	53.0	61.0	68.0	69.4	63.5	68.1	61.8
	AGW [58]	TPAMI'21	47.5	47.7	54.2	63.0	70.1	66.4	70.5	65.9
	NFS [59]	CVPR'21	56.9	55.5	62.8	69.8	80.5	72.1	78.0	69.8
	LbA [60]	ICCV'21	55.4	54.1	58.5	66.3	74.2	67.6	72.4	65.5
	CAJ [1]	ICCV'21	69.9	66.9	76.3	80.4	85.0	79.1	84.8	77.8
	MPANet [40]	CVPR'21	70.6	68.2	76.7	81.0	83.7	80.9	82.8	80.7
	DART [27]	CVPR'22	68.7	66.3	72.5	78.2	83.6	75.7	82.0	73.8
	FMCNet [38]	CVPR'22	66.3	62.5	68.2	74.1	89.1	84.4	88.4	83.9
	MID [61]	AAAI'22	60.3	59.4	64.9	70.1	87.5	84.9	84.3	81.4
	LUPI [62]	ECCV'22	71.1	67.6	82.4	82.7	88.0	82.7	86.8	81.3
	DEEN [63]	CVPR'23	74.7	71.8	80.3	83.3	91.1	85.1	89.5	83.4
	SGIEL [12]	CVPR'23	77.1	72.3	82.1	83.0	92.2	86.6	91.1	85.2
	PartMix [64]	CVPR'23	77.8	74.6	81.5	84.4	85.7	82.3	84.9	82.5
	CAL [65]	ICCV'23	74.7	71.7	79.7	83.7	94.5	88.7	93.6	87.6
	MUN [66]	ICCV'23	76.2	73.8	79.4	82.1	95.2	87.2	91.9	85.0
	SAAI [13]	ICCV'23	75.9	77.0	83.2	88.0	91.1	91.5	92.1	92.0
	FDNM [67]	arXiv'24	77.8	75.1	87.3	89.1	95.5	90.0	94.0	88.7
	LCNL [68]	IJCV'24	70.2	68.0	76.2	80.3	85.6	78.7	84.0	76.9
SSVI-ReID	OTLA [17]	ECCV'22	48.2	43.9	47.4	56.8	49.9	41.8	49.6	42.8
	TAA [19]	TIP'23	48.8	42.3	50.1	56.0	62.2	56.0	63.8	56.5
	DPIS [18]	ICCV'23	58.4	55.6	63.0	70.0	62.3	53.2	61.5	52.7
USVI-ReID	H2H [49]	TIP'21	30.2	29.4	-	-	23.8	18.9	-	-
	OTLA [17]	ECCV'22	29.9	27.1	29.8	38.8	32.9	29.7	32.1	28.6
	ADCA [20]	MM'22	45.5	42.7	50.6	59.1	67.2	64.1	68.5	63.8
	NGLR [69]	MM'23	50.4	47.4	53.5	61.7	85.6	76.7	82.9	75.0
	MBCCM [70]	MM'23	53.1	48.2	55.2	62.0	83.8	77.9	82.8	76.7
	CCLNet [51]	MM'23	54.0	50.2	56.7	65.1	69.9	65.5	70.2	66.7
	PGM [23]	CVPR'23	57.3	51.8	56.2	62.7	69.5	65.4	69.9	65.2
	GUR* [22]	ICCV'23	61.0	57.0	64.2	69.5	73.9	70.2	75.0	69.9
	MMM [54]	ECCV'24	61.6	57.9	64.4	70.4	89.7	80.5	85.8	77.0
	PCLHD	**Ours**	**64.4**	**58.7**	**69.5**	**74.4**	**84.3**	**80.7**	**82.7**	**78.4**
	PCLHD+MMM	**Enhanced**	**65.9**	**61.8**	**70.3**	**74.9**	**89.6**	**83.7**	**87.0**	**80.9**

Table 2: Ablation studies on SYSU-MM01 in all search mode and indoor search mode. "Baseline" means the model trained following PGM [23]. Rank-R accuracy(%) and mAP(%) are reported.

	Component				All Search		Indoor Search	
Index	Baseline	HPCL	DPCL	PCL	Rank-1	mAP	Rank-1	mAP
1	✓				56.3	51.7	60.5	66.2
2	✓		✓		59.1	54.4	63.6	68.8
3	✓	✓			62.1	56.8	65.2	69.8
4	✓	✓	✓		63.7	57.8	67.0	72.6
5	✓	✓	✓	✓	64.4	58.7	69.5	74.4

4.1 Experiment Setting

Dataset. We evaluate our method on two common benchmarks in VI-ReID: **SYSU-MM01** [71] and **RegDB** [72]. SYSU-MM01 is a large-scale public benchmark for the VI-ReID task, which contains 491 identities captured by four RGB cameras and two IR cameras in both outdoor and indoor environments. In this dataset, 22,258 RGB images and 11,909 IR images with 395 identities are collected for training. In the inference stage, the query set consists of 3,803 IR images with 96 identities and the galley set contains 301 randomly selected RGB images. RegDB is collected by an RGB camera and an IR camera, which contains 4,120 RGB images and 4,120 IR images with 412 identities. To be specific, the dataset is randomly divided into two non-overlapping sets: one set is used for training and the other is for testing.

Evaluation Protocols. The experiment follows the standard evaluation settings in VI-ReID, i.e., Cumulative Matching Characteristics (CMC) [73] and mean Average Precision (mAP).

Implementation Details. We adopt the feature extractor in AGW [58], which is initialized with ImageNet-pretrained weights to extract 2048-dimensional features. During the training stage, the

https://doi.org/10.52202/079017-3163

input images are resized to 288×144. We follow augmentations in CAJ [1] for data augmentation. In one batch, we randomly sample 16 pseudo identities, and each pseudo identity samples 16 instances. We set M to be 16 for computational convenience. The number of epochs is 100, in which the first 50 epochs are trained by contrastive loss with the centroid prototype. For the last 50 epochs, we train the model by contrastive loss with both the hard and dynamic prototypes. E_{CPCL} is 50. At the beginning of each epoch, we utilize the DBSCAN [50] algorithm to generate pseudo labels. During the inference stage, we use the momentum encoder ϕ_m to extract features and take the features of the global average pooling layer to calculate cosine similarity for retrieval. The momentum value α and β is set to 0.1 and 0.999, respectively. The temperature hyper-parameter τ is set to 0.05 and the weighting hyper-parameter λ in Eq.(23) is 0.5.

4.2 Results and Analysis

To comprehensively evaluate our method, we compare our method with 18 supervised VI-ReID methods, 3 semi-supervised VI-ReID methods, and 9 unsupervised VI-ReID methods. The comparison results on the SYSU-MM01 and RegDB are reported in Tab. 1.

Comparison with USVI-ReID Methods. As shown in Tab. 1, our method achieves superior performance compared with state-of-the-art USVI-ReID methods. MMM [54] is proposed to establish reliable cross-modality correspondences and is also the current best-performing method. Our method with MMM can achieve 65.9% in Rank-1 and 61.8% in mAP, which surpasses that of MMM by a large margin of 4.3% and 3.9%. Notably, our method even without MMM gains the best performance with 64.4% in Rank-1 and 58.7% in mAP. Although existing USVI-ReID methods mentioned in Tab. 1 have made great progress in the USVI-ReID task, the neglects of divergence and variety hinder their further improvement. They overlook divergence and variety, which often constitutes hard samples. Thus, we propose progressive contrastive learning with hard and dynamic prototypes to mine hard samples, which can guide the model to learn more robust and discriminative features.

Comparison with SSVI-ReID Methods. There are three SSVI-ReID methods proposed to alleviate the problem of labeling cost by using a part of annotations. Remarkably, our method achieves superior performance without any annotations, outperforming all existing SSVI-ReID methods that utilize partial annotations. Moreover, the results suggest that our method can significantly reduce the dependency on manual annotations.

Comparison with SVI-ReID Methods. Surprisingly, our method without annotation outperforms several SVI-ReID methods, e.g., DDAG [39], AGW [58], NFS [59], LbA [60]. This shows the immense competitiveness of PCLHD compared to SVI-ReID methods that rely on complete data annotations. The superior performance of PCLHD mainly benefits from the hard prototype and dynamic prototype contrastive learning. Additionally, we have to acknowledge that a significant disparity still exists between PCLHD and the state-of-the-art fully-supervised results.

4.3 Ablation Study

We conduct ablation studies on the SYSU-MM01 dataset in both all-search and indoor-search modes to show the effectiveness of each component in our method. The results are shown in Tab. 2.

Baseline Settings. We use PGM [23] as our baseline. Although PGM has achieved a promising performance on the USVI-ReID task, the neglect of hard samples hinders its further improvement.

Effectiveness of HPCL. The HPCL is proposed to mine divergence. As shown in Tab. 2, When adding the HPCL on Baseline, the performance improves a large margin of 5.8% in Rank-1 and 5.1% in mAP, respectively. It shows that divergence can be effectively mined using hard prototype contrast learning, facilitating the model to learn more discriminative features.

Effectiveness of DPCL. The DPCL is proposed to mine variety. The results show that Rank-1 accuracy can be improved by 2.8% in Rank-1 and 2.7% in mAP when adding the DPCL on Baseline, which confirms that contrastive learning with dynamic prototype can learn variety.

Effectiveness of PCL. PCL is introduced to smoothly shift the model's attention from commonality to divergence and variety. The results show that Rank-1 accuracy can be improved by about 1% in Rank-1 and mAP compared to adding simultaneously the HPCL and DPCL on the Baseline. This confirms that progressive contrastive learning plays a valuable role in assisting HPCL and DPCL.

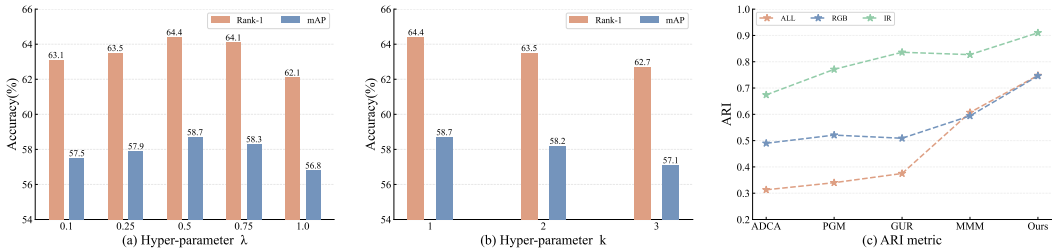

Figure 2: (a) The effect of hyper-parameter λ with different values. (b) The effect of hyper-parameter k with different values. (c) Comparisons with ARI values of different methods.

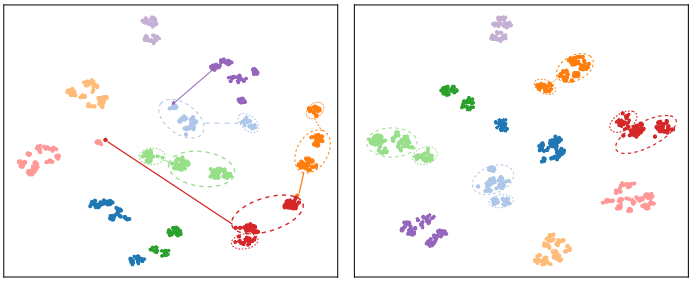

Figure 3: The t-SNE visualization of 10 randomly selected identities. Different color indicates different IDs. Circle means visible features and the pentagram means infrared features.

Surprisingly, contrastive learning with both hard and dynamic prototypes significantly exceeds the baseline by a large margin of 8.1% in Rank-1 and 7.0% in mAP. The HPCL and DPCL can complement each other to learn divergence and variety, which effectively guides the network to learn more robust and discriminative features.

4.4 Further Analysis

Hyper-parameter analysis. Hyper-parameter λ is a weighting parameter to trade-off L_{HPCL} and L_{DPCL}. Fig. 2 (a) presents the results under different values of λ. We can observe that when λ is small, i.e., L_{DPCL} contributes more to the model, the performance degrades. However, when λ is large, i.e., L_{HPCL} contributes heavily to the model, the model both achieves superior performance. Note that when $\lambda = 1$, i.e., the proposed method is trained without DPCL, the performance drops significantly. λ is finally set to 0.5 and our method achieves the best performance of 64.4% in Rank-1. Moreover, we also analyze the effect of the number of hard samples at hard prototype. As shown in Fig. 2 (b), we vary the k from 1 to 3 and keep the other hyper-parameters fixed, which shows that PCLHD achieves the best performance when $k = 1$. Hard samples are distributed in multiple directions, so multiple hard samples cannot be represented by a single prototype. This is why using more hard samples as prototypes leads to a decline in overall performance

The ARI metric. Following MMM [54], we utilize the Adjusted Rand Index (ARI) metric for clustering evaluation. The larger the ARI value, the higher the clustering quality. In Fig. 2 (c), "RGB" and "IR" denote the ARI values of visible and infrared clusterings, which can measure the quality of visible and infrared pseudo-labels. "ALL" means the ARI values of overall clusterings, which can evaluate the reliability of cross-modality correspondences. PCLHD surpasses other methods significantly on all of the mentioned ARI values, which demonstrates PCLHD can effectively mine divergence and variety to improve clustering quality.

Visualization. As shown in Fig. 3, we visualize the t-SNE map of 10 randomly chosen identities from SYSU-MM01. Compared to the baseline, the distribution of the same identity from the same modality is more compact and the distance of the same identity from different modalities is closer together. Moreover, some hard samples in the baseline are incorrectly clustered, while these hard samples are well clustered in our PCLHD, which shows the effectiveness of the proposed PCLHD.

https://doi.org/10.52202/079017-3163

5 Conclusion and Limitation

In this paper, we propose a novel method for USVI-ReID called Progressive Contrastive Learning with hard and dynamic prototype (PCLHD), which learns commonality, divergence and variety. To be specific, we design Hard Prototype Contrastive Learning to mine divergent yet significant information and Dynamic Prototype Contrastive Learning to preserve intrinsic variety features. Furthermore, we introduce a progressive learning strategy to incorporate both HPCL and DPCL into the model. Extensive experiments demonstrate that PCLHD outperforms state-of-the-art USVI-ReID methods.

This work relies on DBSCAN to generate pseudo-labels. However, for extremely large-scale datasets, DBSCAN's performance may be limited, which could affect the overall effectiveness of our approach. To address the limitation, we plan to explore hierarchical clustering in future research to better handle large-scale datasets.

Broader Impacts

This work was developed using publicly available datasets and aims to enhance the capabilities of VI-ReID, which plays a vital role in scenarios where traditional ReID systems fail, such as in low-light or nighttime conditions. VI-ReID offers significant benefits in improving security and surveillance by enabling more reliable identification across varying environmental conditions. Importantly, this work raises no ethical, safety, or environmental concerns, and no harm was inflicted on living beings during the research. However, we acknowledge the risk of misuse, particularly privacy invasion if used to track individuals in public spaces without appropriate regulation. While VI-ReID does not directly identify specific individuals, its unauthorized deployment could still result in significant privacy violations. Therefore, public surveillance systems using VI-ReID should be controlled by authorized entities, ensuring proper regulatory frameworks and adherence to ethical standards.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 62176224, 62176092, 62222602, 62306165, 62106075, 62476090), Natural Science Foundation of Shanghai (23ZR1420400), Natural Science Foundation of Chongqing (CSTB2023NSCQ-JQX0007), China Computer Federation (CCF) Lenovo Blue Ocean Research Fund, China Academy of Railway Sciences No. 2023YJ357.

References

[1] Mang Ye, Weijian Ruan, Bo Du, and Mike Zheng Shou. Channel augmented joint learning for visible-infrared recognition. In *ICCV*, pages 13547–13556, 2021.

[2] Yunpeng Gong, Zhun Zhong, Zhiming Luo, Yansong Qu, Rongrong Ji, and Min Jiang. Cross-modality perturbation synergy attack for person re-identification. *arXiv preprint arXiv:2401.10090*, 2024.

[3] Jiangming Shi, Xiangbo Yin, Demao Zhang, and Yanyun Qu. Visible embraces infrared: Cross-modality person re-identification with single-modality supervision. In *China Automation Congress*, pages 4781–4787, 2023.

[4] Jiaxu Leng, Zhanjie Wu, Mingpi Tan, Yiran Liu, Ji Gan, Haosheng Chen, and Xinbo Gao. Beyond euclidean: Dual-space representation learning for weakly supervised video violence detection. *arXiv preprint arXiv:2409.19252*, 2024.

[5] Yuanmin Tang, Jing Yu, Keke Gai, Jiamin Zhuang, Gang Xiong, Yue Hu, and Qi Wu. Context-i2w: Mapping images to context-dependent words for accurate zero-shot composed image retrieval. In *AAAI*, pages 5180–5188, 2024.

[6] Yuanmin Tang, Jing Yu, Keke Gai, Jiamin Zhuang, Gaopeng Gou, Gang Xiong, and Qi Wu. Denoise-i2w: Mapping images to denoising words for accurate zero-shot composed image retrieval. *arXiv preprint arXiv:2410.17393*, 2024.

[7] Junjie Zhou, Zheng Liu, Shitao Xiao, Bo Zhao, and Yongping Xiong. Vista: Visualized text embedding for universal multi-modal retrieval. In *ACL*, pages 3185–3200, 2024.

[8] Yachao Zhang, Runze Hu, Ronghui Li, Yanyun Qu, Yuan Xie, and Xiu Li. Cross-modal match for language conditioned 3d object grounding. In *AAAI*, volume 38, pages 7359–7367, 2024.

[9] Yachao Zhang, Yuan Xie, Cuihua Li, Zongze Wu, and Yanyun Qu. Learning all-in collaborative multiview binary representation for clustering. *IEEE TNNLS*, 35(3):4260–4273, 2022.

[10] Ling Lin, Hao Liu, Jinqiao Liang, Zhendong Li, Jiao Feng, and Hu Han. Consensus-agent deep reinforcement learning for face aging. *IEEE Transactions on Image Processing*, 2024.

[11] Ling Lin, Tao Wang, Hao Liu, Congcong Zhu, and Jingrun Chen. Toward quantifiable face age transformation under attribute unbias. *IEEE TCSVT*, 2024.

[12] Jiawei Feng, Ancong Wu, and Wei-Shi Zheng. Shape-erased feature learning for visible-infrared person re-identification. In *CVPR*, pages 22752–22761, 2023.

[13] Xingye Fang, Yang Yang, and Ying Fu. Visible-infrared person re-identification via semantic alignment and affinity inference. In *ICCV*, pages 11270–11279, 2023.

[14] Rui Sun, Long Chen, Lei Zhang, Ruirui Xie, and Jun Gao. Robust visible-infrared person re-identification based on polymorphic mask and wavelet graph convolutional network. *IEEE TIFS*, 2024.

[15] Xiangbo Yin, Jiangming Shi, Yachao Zhang, Yang Lu, Zhizhong Zhang, Yuan Xie, and Yanyun Qu. Robust pseudo-label learning with neighbor relation for unsupervised visible-infrared person re-identification. In *ACM MM*, 2024.

[16] Bo Pang, Deming Zhai, Junjun Jiang, and Xianming Liu. Fully unsupervised person re-identification via selective contrastive learning. *ACM TOMM*, 18(2):1–15, 2022.

[17] 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.

[18] Jiangming Shi, Yachao Zhang, Xiangbo Yin, Yuan Xie, Zhizhong Zhang, Jianping Fan, Zhongchao Shi, and Yanyun Qu. Dual pseudo-labels interactive self-training for semi-supervised visible-infrared person re-identification. In *ICCV*, pages 11218–11228, 2023.

[19] Bin Yang, Jun Chen, Xianzheng Ma, and Mang Ye. Translation, association and augmentation: Learning cross-modality re-identification from single-modality annotation. *IEEE TIP*, 32:5099–5113, 2023.

[20] 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.

[21] Guoqing Zhang, Hongwei Zhang, Weisi Lin, Arun Kumar Chandran, and Xuan Jing. Camera contrast learning for unsupervised person re-identification. *IEEE TCSVT*, 33(8):4096–4107, 2023.

[22] Bin Yang, Jun Chen, and Mang Ye. Towards grand unified representation learning for unsupervised visible-infrared person re-identification. In *ICCV*, pages 11069–11079, 2023.

[23] Zesen Wu and Mang Ye. Unsupervised visible-infrared person re-identification via progressive graph matching and alternate learning. In *CVPR*, pages 9548–9558, 2023.

[24] Lei Tan, Jiaer Xia, Wenfeng Liu, Pingyang Dai, Yongjian Wu, and Liujuan Cao. Occluded person re-identification via saliency-guided patch transfer. In *AAAI*, volume 38, pages 5070–5078, 2024.

[25] Lei Tan, Pingyang Dai, Jie Chen, Liujuan Cao, Yongjian Wu, and Rongrong Ji. Partformer: Awakening latent diverse representation from vision transformer for object re-identification. *arXiv preprint arXiv:2408.16684*, 2024.

[26] Kunlun Xu, Xu Zou, Yuxin Peng, and Jiahuan Zhou. Distribution-aware knowledge prototyping for non-exemplar lifelong person re-identification. In *CVPR*, pages 16604–16613, 2024.

[27] 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 14288–14297, 2022.

[28] Pingping Zhang, Yuhao Wang, Yang Liu, Zhengzheng Tu, and Huchuan Lu. Magic tokens: Select diverse tokens for multi-modal object re-identification. In *CVPR*, pages 17117–17126, 2024.

[29] Yuhao Wang, Xuehu Liu, Pingping Zhang, Hu Lu, Zhengzheng Tu, and Huchuan Lu. Top-reid: Multi-spectral object re-identification with token permutation. In *AAAI*, pages 5758–5766, 2024.

 https://doi.org/10.52202/079017-3163

[30] Jialong Zuo, Hanyu Zhou, Ying Nie, Feng Zhang, Tianyu Guo, Nong Sang, Yunhe Wang, and Changxin Gao. Ufinebench: Towards text-based person retrieval with ultra-fine granularity. In *CVPR*, pages 22010–22019, 2024.

[31] Chenyang Yu, Xuehu Liu, Yingquan Wang, Pingping Zhang, and Huchuan Lu. Tf-clip: Learning text-free clip for video-based person re-identification. In *AAAI*, pages 6764–6772, 2024.

[32] Yukang Zhang, Yan Yan, Jie Li, and Hanzi Wang. Mrcn: A novel modality restitution and compensation network for visible-infrared person re-identification. In *AAAI*, volume 37, pages 3498–3506, 2023.

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

[34] 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 3622–3631, 2019.

[35] 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.

[36] Yukang Zhang, Yan Yan, Yang Lu, and Hanzi Wang. Towards a unified middle modality learning for visible-infrared person re-identification. In *ACM MM*, pages 788–796, 2021.

[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] 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 7339–7348, 2022.

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

[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.

[41] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. In *ICLR*, 2020.

[42] Yixiao Ge, Feng Zhu, Dapeng Chen, Rui Zhao, and Hongsheng Li. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. In *NeurIPS*, 2020.

[43] Neng Dong, Liyan Zhang, Shuanglin Yan, Hao Tang, and Jinhui Tang. Erasing, transforming, and noising defense network for occluded person re-identification. *IEEE TCSVT*, pages 1–1, 2023.

[44] Neng Dong, Shuanglin Yan, Hao Tang, Jinhui Tang, and Liyan Zhang. Multi-view information integration and propagation for occluded person re-identification. *Information Fusion*, 104:102201, 2024.

[45] Yoonki Cho, Woo Jae Kim, Seunghoon Hong, and Sung-Eui Yoon. Part-based pseudo label refinement for unsupervised person re-identification. In *CVPR*, pages 7298–7308, 2022.

[46] Zuozhuo Dai, Guangyuan Wang, Weihao Yuan, Siyu Zhu, and Ping Tan. Cluster contrast for unsupervised person re-identification. In *ACCV*, pages 319–337, 2022.

[47] Chang Zou, Zeqi Chen, Zhichao Cui, Yuehu Liu, and Chi Zhang. Discrepant and multi-instance proxies for unsupervised person re-identification. In *ICCV*, pages 11058–11068, 2023.

[48] Yuhang Wu, Tengteng Huang, Haotian Yao, Chi Zhang, Yuanjie Shao, Chuchu Han, Changxin Gao, and Nong Sang. Multi-centroid representation network for domain adaptive person re-id. In *AAAI*, pages 2750–2758, 2022.

[49] 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.

[50] 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.

[51] Zhong Chen, Zhizhong Zhang, Xin Tan, Yanyun Qu, and Yuan Xie. Unveiling the power of clip in unsupervised visible-infrared person re-identification. In *ACM MM*, pages 3667–3675, 2023.

[52] Mouxing Yang, Yunfan Li, Zhenyu Huang, Zitao Liu, Peng Hu, and Xi Peng. Partially view-aligned representation learning with noise-robust contrastive loss. In *CVPR*, pages 1134–1143, 2021.

[53] Mouxing Yang, Yunfan Li, Peng Hu, Jinfeng Bai, Jiancheng Lv, and Xi Peng. Robust multi-view clustering with incomplete information. *IEEE TPAMI*, 45(1):1055–1069, 2023.

[54] Jiangming Shi, Xiangbo Yin, Yeyun Chen, Yachao Zhang, Zhizhong Zhang, Yuan Xie, and Yanyun Qu. Multi-memory matching for unsupervised visible-infrared person re-identification. In *ECCV*, page 456–474, 2024.

[55] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *ICCV*, pages 9729–9738, 2020.

[56] Lei Tan, Pingyang Dai, Rongrong Ji, and Yongjian Wu. Dynamic prototype mask for occluded person re-identification. In *ACM MM*, pages 531–540, 2022.

[57] Yunpeng Gong, Liqing Huang, and Lifei Chen. Person re-identification method based on color attack and joint defence. In *CVPRW*, pages 4312–4321, 2022.

[58] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C. H. Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE TPAMI*, pages 2872–2893, 2022.

[59] 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.

[60] Hyunjong Park, Sanghoon Lee, Junghyup Lee, and Bumsub Ham. Learning by aligning: Visible-infrared person re-identification using cross-modal correspondences. In *ICCV*, pages 12026–12035, 2021.

[61] Zhipeng Huang, Jiawei Liu, Liang Li, Kecheng Zheng, and Zheng-Jun Zha. Modality-adaptive mixup and invariant decomposition for rgb-infrared person re-identification. In *AAAI*, pages 1034–1042, 2022.

[62] Mahdi Alehdaghi, Arthur Josi, Rafael M. O. Cruz, and Eric Granger. Visible-infrared person re-identification using privileged intermediate information. In *ECCV*, pages 720–737, 2022.

[63] 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.

[64] Minsu Kim, Seungryong Kim, Jungin Park, Seongheon Park, and Kwanghoon Sohn. Partmix: Regularization strategy to learn part discovery for visible-infrared person re-identification. In *CVPR*, pages 18621–18632, 2023.

[65] Jianbing Wu, Hong Liu, Yuxin Su, Wei Shi, and Hao Tang. Learning concordant attention via target-aware alignment for visible-infrared person re-identification. In *ICCV*, pages 11122–11131, 2023.

[66] Hao Yu, Xu Cheng, Wei Peng, Weihao Liu, and Guoying Zhao. Modality unifying network for visible-infrared person re-identification. In *ICCV*, pages 11185–11195, 2023.

[67] Yukang Zhang, Yang Lu, Yan Yan, Hanzi Wang, and Xuelong Li. Frequency domain nuances mining for visible-infrared person re-identification. *arXiv preprint arXiv:2401.02162*, 2024.

[68] Mouxing Yang, Zhenyu Huang, and Xi Peng. Robust object re-identification with coupled noisy labels. *IJCV*, pages 1–19, 2024.

[69] De Cheng, Xiaojian Huang, Nannan Wang, Lingfeng He, Zhihui Li, and Xinbo Gao. Unsupervised visible-infrared person reid by collaborative learning with neighbor-guided label refinement. In *ACM MM*, pages 7085–7093, 2023.

[70] Lingfeng He, Nannan Wang, Shizhou Zhang, Zhen Wang, Xinbo Gao, et al. Efficient bilateral cross-modality cluster matching for unsupervised visible-infrared person reid. In *ACM MM*, pages 1325–1333, 2023.

[71] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. Rgb-infrared cross-modality person re-identification. In *ICCV*, pages 5390–5399, 2017.

[72] 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*, 17(3):605, 2017.

[73] Mang Ye, Zheng Wang, Xiangyuan Lan, and Pong C. Yuen. Visible thermal person re-identification via dual-constrained top-ranking. In *IJCAI*, pages 1092–1099, 2018.

 https://doi.org/10.52202/079017-3163

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: This work relies on DBSCAN to generate pseudo-labels. However, for extremely large-scale datasets, DBSCAN's performance may be limited, which could affect the overall effectiveness of our approach. To address the limitation, we plan to explore hierarchical clustering in future research to better handle large-scale datasets.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: For each theoretical result, the paper provide the full set of assumptions and a complete (and correct) proof.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims. Our code will be released after the acceptance of our paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

https://doi.org/10.52202/079017-3163

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The paper provide open access to the code, with sufficient instructions to faithfully reproduce the main experimental results.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specify all the training and test details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Error bars are not reported because it would be too computationally expensive.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper provide sufficient information on the computer resources.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper was developed using publicly available infrared-visible ReID datasets and aims to enhance the capabilities of visible-infrared ReID, which plays a vital role in scenarios where traditional ReID systems fail, such as in low-light or nighttime conditions. This technology offers significant benefits in improving security and surveillance by enabling more reliable identification across varying environmental conditions. Importantly, our

https://doi.org/10.52202/079017-3163

research raises no ethical, safety, or environmental concerns, and no harm was inflicted on living beings during the research. However, we acknowledge the risk of misuse, particularly privacy invasion if used to track individuals in public spaces without appropriate regulation. While ReID technology does not directly identify specific individuals, its unauthorized deployment could still result in significant privacy violations. Therefore, public surveillance systems using ReID should be controlled by authorized entities, ensuring proper regulatory frameworks, transparency, and adherence to ethical standards.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: They are properly credited and respected.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [No]

    Justification: The datasets used in this paper, SYSU-MM01 and RegDB, are publicly available and widely used in research. These datasets were collected by their original creators and made accessible for research purposes.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

https://doi.org/10.52202/079017-3163

Answer: [No]

Justification: Since we are using datasets that are already publicly available and have beenextensively used in previous research, and given that the content does not involve sensitivepersonal information, this study did not undergo an independent IRB review.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.