Cross-modality Person Re-identification Based on Unsupervised Learning

Xiangbo Yin[1], **Zhuo Chen**[1], **YiJing Zheng**[2], **Hualong Ke**[2], **Yilin Cai**[1]

[1]School of Informatics, [2]Artificial Intelligence
{23020231154251, 23020231154136, 23320231154462, 23020231154197, 23320231154441}@stu.xmu.edu.cn

Abstract

Person re-identification (ReID) has gained a lot of attention in recent years due to its application in video surveillance and security. Traditional person re-identification aims to match the same pedestrians under different visible cameras. However, in poor light environments, visible cameras may not work, hindering the application of pedestrian re-identification in real-life scenarios. To overcome this challenge, the introduction of infrared images as complementary information can enhance the accuracy and robustness of person re-identification by utilizing the thermal energy distribution present in infrared images, which promotes the development of visible-infrared person re-identification (VI-ReID). However, there are significant modality differences between visible and infrared images, making direct visible-infrared image person re-identification a challenging task. In practical applications, obtaining paired annotations for visible-infrared image pairs is expensive and time-consuming, making it difficult to acquire large-scale annotated datasets. To address the aforementioned issues, we propose an unsupervised learning framework for VI-ReID, which contains Modal-specific Cluster Contrast (MSCC) module, Modal-invariant Cluster Contrast (MICC) module, and Prototype Similarity Association (PSA) module. These modules are used to learn modal-specific information, explore modal-invariant information, and establish cross-modal association, respectively. Comprehensive experimental results demonstrate the effectiveness of our proposed approach.

Introduction

Person re-identification (ReID) targets matching the same person across different cameras(Ge et al. 2021; Guo et al. 2019). In recent years, ReID has gained widespread attention due to its application in the field of intelligent monitoring(Wang et al. 2017). However, the previously widely used technology of collecting images using visible light cameras cannot accurately work in low-light scenarios such as nighttime, which greatly limits the performance of intelligent monitoring systems that only use single-modal ReID(Wang et al. 2019b). To address this issue, researchers have proposed cross-modal visible informed person re-identification (VI-ReID), which involves cross-modal fusion recognition of images captured by infrared and visible light cameras(Wu et al. 2017b).

However, on this basis, new issues have emerged in the study. Training VI-ReID models in a supervised manual requires a substantial amount of cross-modal identity annotations, making it expendable and limiting its practical application in real-world scenarios.

Motivated by the aforementioned factors, this article delves into a complex unsupervised learning challenge known as USL-VI-ReID. The objective of this task is to extract modality-invariant knowledge from unlabeled datasets of visible and infrared images, enabling the identification of individuals captured by both types of cameras. We propose an efficient collaborative unsupervised learning framework for VI-ReID, as shown in Fig. 1. The proposed method consists of three main modules: 1) Modal-specific Cluster Contrast learning (MSCC), 2) Modal-invariant Cluster Contrast learning (MICC), and 3) Prototype Similarity Association (PSA). To be specific, the function of each module is as follows:

- **Modal-specific Cluster Contrast learning.** As shown in Fig. 1(a), We first employ a dual-stream network to extract the modal-specific features of visible and infrared images, respectively. Afterward, we cluster the features by DBSCAN and compute the prototypes of each cluster, which are used to initialize modal-specific memories. In each iteration, the modality-specific shallow layers are updated by the ClusterNCE loss (Dai et al. 2021).

- **Modal-invariant Cluster Contrast learning.** Unlike the previous methods, we not only cluster intra-modal features but also inter-modal features, which helps the model better learn modal-invariant information. As shown in Fig. 1(b), different from MSCC, MICC regards clusters containing both visible and infrared features as available clusters and discards clusters with only single-modal features.

- **Prototype Similarity Association.** Based on the similarity between visible and infrared prototypes, we use triplet loss to minimize the distance between the positive-pair cluster centers while increasing that between the negative-pair cluster centers, thereby further exploring the correlation between visible and infrared features (See Fig. 1(c)).

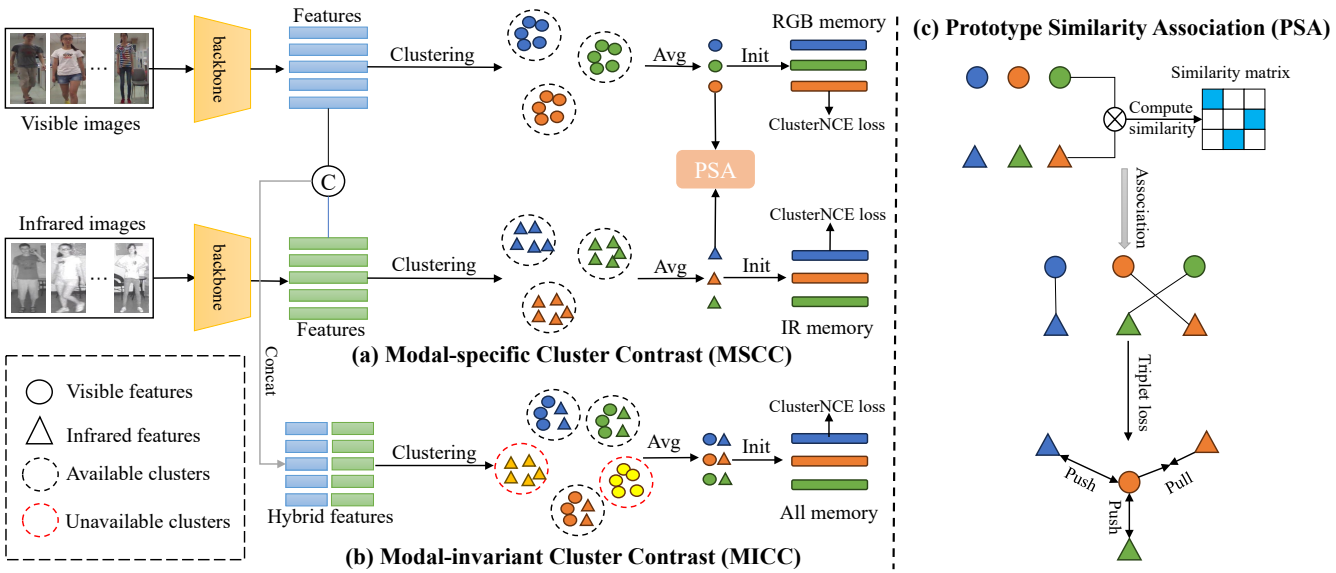

Figure 1: The framework of our method.

Related Work

Supervised Visible-Infrared Person ReID.
Visible-infrared person ReID (VI-ReID) poses a formidable cross-modality person recognition challenge due to the substantial modality gap between visible and infrared intra-class person images (Dai et al. 2018; Hao et al. 2021; Tian et al. 2021; Ye et al. 2020). Various techniques have been developed to address this challenge. However, most supervised VI-ReID methods, like those proposed by (Wu et al. 2017c; Wang et al. 2019a; Wu et al. 2021), heavily rely on abundant cross-modality identity annotations, limiting their versatility in diverse scenarios.

Unsupervised Single-Modality Person ReID.
Traditional unsupervised person ReID methods can be broadly categorized into two groups: fully unsupervised learning (USL) methods and unsupervised domain adaptation (UDA) methods, depending on whether or not they utilize labeled source domain datasets for model training. In the case of UDA, some approaches (Chen et al. 2021; Chen, Zhu, and Gong 2019; Wei et al. 2018) employ generative techniques to transfer knowledge from the source domain to the target domain. Unsupervised Learning (USL) offers a more challenging yet pragmatic approach for real-world scenarios, as it circumvents the need for labeled source domain datasets. In recent years, many USL methods (Dai et al. 2022a; Wang et al. 2021; Zhang et al. 2022, 2021) have typically leveraged pseudo-labels generated by clustering methods to optimize the model. These techniques have further refined the quality of pseudo-labels.

Unsupervised Visible-Infrared Person ReID.
Unsupervised Visible-Infrared Person ReID (USL-VI-ReID) problem has two challenges. First, different from single-modality person ReID, USL-VI-ReID has a large cross-modality discrepancy, which results in large intra-class variations and makes it difficult to generate reliable cross-modality labels. Second, there are no available cross-modality (visible-infrared) identity labels in USL-VI-ReID, leading to the difficulty in directly learning modality-invariant feature representations. H2H (Liang et al. 2021) started the first attempt by designing a two-stage method to solve the USL-VI-ReID task, including homogeneous learning and heterogeneous learning. OTLA (Wang et al. 2022) tries to assign the infrared images to the pseudo-visible labels based on the optimal transport strategy. However, these methods require extra RGB datasets for pre-training, making the method less scalable in real-world deployments.

Method

Formulation and Overview

Let $X = \{V, R\}$ denote an unlabeled VI-ReID dataset, where $V = \{x_i^v\}_{i=1}^{N_v}$ and $R = \{x_i^r\}_{i=1}^{N_r}$ denote N_v visible images and N_r infrared images from two modalities, respectively. In the USL-VI-ReID task, our goal is to train a deep neural network $f_\theta(\cdot)$ to project an image x_i from the dataset X into an embedding space F_θ and obtain a d-dimensional modality-invariant representation $u_i = f_\theta(x_i) \in \mathbb{R}^d$.

We first utilize DBSCAN(Ester et al. 1996b) algorithm to obtain pseudo labels $Y_V = \{y_i^v\}_{i=1}^{K_v}$ and $\mathrm{Y_R} = \{y_i^r\}_{i=1}^{K_r}$ for unlabeled samples from two modalities, where Y_v and Y_r denote the number of clustering of Visible and infrared modalities, respectively. A two-stream encoder with modality-specific shallow layers and modality-shared deep layers is used to extract features.

Modal-specific Cluster Contrast

Different statistical properties of multi-modal data hint it is rather difficult to fuse different modalities in the data space. In light of this, we design a modality-specific cluster contrast

learning module to transform the data into low-dimensional latent space. This module performs the main task of feature learning. At the beginning of each training epoch, we initialize two modality-specific memory banks M^{υ} and M^{r} by the cluster centroids $\{\phi_i^{\upsilon}\}_{i=1}^{K_v}$ and $\{\phi_i^{r}\}_{i=1}^{K_r}$ respectively, to store the representations of clusters from each modality and update them with a momentum strategy (Chen, Lagadec, and Bremond 2021; Dai et al. 2022b), where K_{υ} and K_r are the numbers of grouped clusters in these two modalities. This training process enables the encoder to extract expressive features for each modality and generate high-quality pseudo labels. This process can be written as:

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

$$\phi_l^v = \frac{1}{\left|\mathrm{H}_l^v\right|} \sum_{u_n^v \in \mathrm{H}_l^v} u_n^v, \tag{2}$$

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

Modality-specific Memory Updating. During training, we sample P person identities and Z instances for each identity from each modality training set. Then, we obtain a total number of $3P \times Z$ query images including infrared, visible, and augmented visible person images in a batch. We update the two modality-specific memories by a momentum updating strategy:

$$\phi_k^{i(\delta)} \leftarrow \beta \phi_k^{i(\delta-1)} + (1-\beta) q_i, \tag{3}$$

$$\phi_l^{v(\delta)} \leftarrow \beta \phi_l^{v(\delta-1)} + (1-\beta) q_v, \tag{4}$$

$$\phi_l^{\upsilon(\delta)} \leftarrow \beta \phi_l^{\upsilon(\delta-1)} + (1-\beta) q_{\upsilon a}, \tag{5}$$

where $q_{\upsilon a}$ is the augmented query instance features. β is the momentum updating factor. δ is the iteration number.

Joint Learning Loss Function. In each iteration, the modality-specific shallow layers and shared layers are jointly updated by three types of ClusterNCE(Dai et al. 2022c) loss, including infrared, visible, and augmented visible loss by the following equations:

$$L_{q_i} = -\log \frac{\exp\left(q_i \cdot \phi_+^i / \tau\right)}{\sum_{k=0}^{K} \exp\left(q_i \cdot \phi_k^i / \tau\right)}, \tag{6}$$

$$L_{q_v} = -\log \frac{\exp\left(q_v \cdot \phi_+^v / \tau\right)}{\sum_{l=0}^{L} \exp\left(q_v \cdot \phi_k^v / \tau\right)}, \tag{7}$$

$$L_{q_{va}} = -\log \frac{\exp\left(q_{va} \cdot \phi_+^{\upsilon} / \tau\right)}{\sum_{l=0}^{L} \exp\left(q_{va} \cdot \phi_k^{\upsilon} / \tau\right)}, \tag{8}$$

where ϕ_+ is the positive representation vector of the cluster corresponding to the pseudo label of the query and the τ is a temperature hyper-parameter following Cluster Contrast.

MSCC Loss Function. Certainly, three types of ClusterNCE loss function are designed to learn discriminative representation:

$$L_{MSCC} = L_{q_i} + L_{q_v} + L_{q_{va}}. \tag{9}$$

The loss value is low when q is close to its positive cluster representation and dissimilar to all other cluster features. This process is equivalent to two modality-specific softmax-based classifiers that try to classify q_i as ϕ_+^i, and q_{υ} together with q_{va} as q_{va} as ϕ_+^v. q_{va} is the query of channel augmented features for learning color-invariant information, and thus feature encoders have a certain modality generalization ability with the help of joint augmented learning.

Modal-invariant Cluster Contrast

Through the Modal-specific Cluster Contrast(MSCC) introduced above, the model has learned the specific characteristics of a specific mode and has the generalization ability of a single mode. However, the model cannot handle the correspondence between different modalities well, making it difficult to truly achieve cross-modality pedestrian re-identification. In order to better establish cross-modality connections, we have designed a Modal-invariant Cluster Contrast(MICC) module to better capture invariant features between modalities, thereby having stronger ability for cross-modality pedestrian re-identification.

We put the features of visible and infrared features together forming a hybrid set $\{u_1^v, ..., u_{N_v}^v, u_1^i, ..., u_{N_r}^i\}$, and use the classic DBSCAN algorithm to cluster. Clusters containing both visible and infrared features are regarded as available clusters, and those with only single modal features are discarded. This operation helps the network better capture cross modal features and establish relationships between them. We will average the visible and infrared features belonging to the same cluster to obtain the center feature of the cross-modality cluster. The specific operation can be described as

$$\phi_g^{vi} = \frac{1}{\left|\mathrm{H}_g^{vi}\right|} \sum_{u_n^{vi} \in \mathrm{H}_g^{vi}} u_n^{vi}. \tag{10}$$

where $\mathrm{H}_g^{\mathrm{vi}(v)}$ denotes the g-th cross-modality cluster and $|\cdot|$ indicates the number of instances per cluster.

Memory Aggregation. We aggregate the selected memories using a momentum updating strategy by

$$\phi_g^{vi(\delta)} \leftarrow \beta \phi_g^{vi(\delta-1)} + (1-\beta) q_i, \tag{11}$$

$$\phi_g^{vi(\delta)} \leftarrow \beta \phi_g^{vi(\delta-1)} + (1-\beta) q_v, \tag{12}$$

Both visible and infrared features are used to update their common cross-modality memory.

MSCC Loss Function. We also use ClusterNCE loss to promote cross- modality feature learning by the following equation:

Table 1: Comparison with the state-of-the-art VI-ReID methods on SYSU-MM01 dataset. It contains supervised and unsupervised ReID methods. Rank-k accuracy (%), mAP(%) are reported.

	SYSU-MM01 Settings		All-search				Indoor Search			
	Methods	Venue	r1(%)	r10(%)	r20%	mAP(%)	r1(%)	r10(%)	r20%	mAP(%)
Supervised	AlignGAN	ICCV-19	42.40	85.0	93.7	40.70	45.90	87.60	94.40	54.30
	cm-SSFT	TPAMI-20	47.70	-	-	54.10	-	-	-	-
	AGW	CVPR-20	47.50	84.39	92.14	47.65	54.17	91.14	95.98	62.97
	DDAG	ECCV-20	54.75	90.39	95.81	53.02	61.02	94.06	98.41	67.98
	CA	ICCV-21	69.88	95.71	98.46	66.89	76.26	97.88	99.49	80.37
Unsupervised	CAP	AAAI-21	16.82	47.60	61.42	15.71	24.57	57.93	72.74	30.74
	ICE	ICCV-21	20.54	57.5	70.89	20.39	29.81	69.41	82.66	38.35
	PPLR	CVPR-22	12.58	47.43	62.69	12.78	13.65	52.66	70.28	22.19
	ISE	CVPR-22	20.01	57.45	72.50	18.93	14.22	58.33	75.32	24.62
	H2H	TIP-21	30.15	65.92	77.32	29.40	-	-	-	-
	ADCA	MM-22	45.51	85.29	93.16	42.73	50.60	89.66	96.15	59.11
	Ours	-	**53.3**	**90.8**	**96.5**	**50.2**	**57.3**	**93.7**	**97.6**	**64.7**

$$L_{MICC} = -\log \frac{\exp\left(q_v \cdot \phi_+^{vi}/\tau\right)}{\sum_{g=0}^{G} \exp\left(q_v \cdot \phi_g^{vi}/\tau\right)} - \log \frac{\exp\left(q_i \cdot \phi_+^{vi}/\tau\right)}{\sum_{g=0}^{G} \exp\left(q_i \cdot \phi_g^{vi}/\tau\right)}. \tag{13}$$

Prototype Similarity Association

Thanks to MICC and MSCC, the model is able to learn both modal-specific and modal-invariant features simultaneously, and has a certain ability to cross-modality recognition. In order to further explore the correlation between visual features and infrared features, we propose the Prototype Similarity Association (PSA) module. Firstly, we calculate the cosine similarity between the center of the visual feature cluster and the center of the infrared feature cluster, and construct a similarity matrix to measure the correlation between cross-modality features.

$$\text{Similarity } \left(\phi^v, \phi^i\right) = \frac{\phi^v \cdot \phi^i}{||\phi^v|| \times ||\phi^i||}. \tag{14}$$

where ϕ^v and ϕ^i denote the center of the visual feature cluster and the infrared feature cluster respectively.

PSA Loss function. For a visual feature center, it forms a positive sample pair with the infrared feature center that has the highest similarity with it, and a negative sample pair with the other infrared feature centers. We assign new pseudo labels $< y^v, y^i >$ to positive sample pairs. Triplet Loss is used to bring together high similarity cross-modality features while distancing low similarity cross-modality features by the following equation:

$$L_{PSA}^v = max(d(\phi_a^v, \phi_p^i) - d(\phi_a^v, \phi_n^i) + m, 0), \tag{15}$$

$$L_{PSA}^i = max(d(\phi_a^i, \phi_p^v) - d(\phi_a^i, \phi_n^v) + m, 0), \tag{16}$$

where $d(\cdot,\cdot)$ denotes the Euclidean Distance. ϕ^a, ϕ^p and ϕ^n represent the features cluster centers of anchor, positive, and negative respectively. m is the margin controlling how much higher the distance with the negative cluster is than with the positive cluster. The total PSA loss can be described as:

$$L_{PSA} = L_{PSA}^v + L_{PSA}^i. \tag{17}$$

Overall Loss Function. In summary, the final loss function can be defined as:

$$L = L_{MSCC} + L_{MICC} + L_{PSA}. \tag{18}$$

Experiments

Expeimental Setting

Dataset. We utilized two publicly available datasets, namely RegDB (Park 2017) and SYSU-MM01 (Wu et al. 2017a) datasets, to evaluate our cross-modality person re-identification approach. The SYSU-MM01 dataset comprises cross-modality person images captured by two near-infrared cameras and four visible cameras. It encompasses 395 training identities, containing a total of 22,258 visible images and 11,909 near-infrared images, captured in various indoor and outdoor environments. Our evaluation encompasses both all-search and indoor-search modes.The RegDB dataset is captured using a system that consists of two aligned cameras, one thermal and one visible. It consists of 412 unique identities. To assess our method, we conduct evaluations in two test modes: thermal to visible and visible to thermal. We adhere strictly to established methodologies, performing ten trials of gallery set selection, and subsequently compute the average performance.

Evaluation Metrics. Consistent with previous studies (Mang Ye and Yuen 2018), we employ Mean Average Precision (mAP) and Cumulative Matching Characteristics (CMC) as the evaluation metrics for our analysis.

Implentation Details. We have implemented our proposed framework using PyTorch. The configuration of the two shallow layers adheres to the approach introduced in AGW (Ye et al. 2022). For the shared layers, we utilize ResNet50, which is initialized with pre-trained weights from ImageNet. During the testing phase, we extract features from the pooling layer of GeM (Radenovic, Tolias, and Chum 2019)

	RegDB Settings		Visible to Thermal				Thermal to Visible			
	Methods	Venue	r1(%)	r10(%)	r20%	mAP(%)	r1(%)	r10(%)	r20%	mAP(%)
Supervised	AlignGAN	ICCV-19	57.9	-	-	53.6	56.3	-	-	53.4
	cm-SSFT	CVPR-20	72.3	-	-	72.9	71.0	-	-	71.7
	AGW	TPAMI-21	70.05	86.21	91.15	66.37	70.49	87.21	91.84	65.90
	DDAG	ECCV-20	69.34	86.19	91.49	63.46	68.06	85.15	90.31	61.80
	CA	ICCV-21	85.03	95.49	97.54	79.14	84.75	95.33	97.51	77.82
Unsupervised	CAP	AAAI-21	9.71	19.27	25.6	11.56	10.21	19.91	26.38	11.34
	ICE	ICCV-21	12.98	25.87	34.4	15.64	12.18	25.67	34.9	14.82
	PPLR	CVPR-22	8.93	20.87	27.91	11.14	8.11	20.29	28.79	9.07
	ISE	CVPR-22	16.12	23.30	28.93	16.99	10.83	18.64	27.09	13.66
	H2H	TIP-21	23.81	45.31	54.00	18.87	-	-	-	-
	ADCA	MM-22	67.20	82.02	87.44	64.05	68.48	83.21	88.00	63.81
	Ours	-	**74.4**	**87.7**	**91.5**	**70.3**	**73.5**	86.3	89.8	**70.1**

and calculate cosine similarity. At the start of each training epoch, DBSCAN (Ester et al. 1996a) is employed to generate pseudo labels independently for each modality.

During the training process, we select 16 person identities and 16 instances for each identity from the training sets of each modality. To enhance the training, we apply random horizontal flipping, random erasing, and random cropping techniques to images with a size of 288×144. In the case of the augmented visible stream, we employ the random Channel Augmentation (CA) method. The model is trained using the Adam optimizer. Initially, the learning rate is set to 3.5e-4 and then reduced to 1/10 of its previous value every 20 epochs. We train the model for a total of 100 epochs, where the first 50 epochs are dedicated to pre-training the MSCC framework, and the MICC and PSA are performed in the last 50 epochs.

Results and Analysis

Comparision with SVI-ReID Methods. As shown in Fig. 1 Our method outperforms several supervised methods. This is an encouraging outcome, demonstrating the potential of unsupervised cross-modality ReID in approximating the effectiveness of supervised VI-ReID.

The significant improvements observed in our method can be attributed to its insightful design for USL-VI-ReID, offering three key advantages. Firstly, we do not require any additional labeled data, making our proposed framework more practical for deployment. Secondly, our solution is simple, efficient, and easy to implement. We anticipate that incorporating advanced contrastive learning techniques would further enhance performance. Lastly, the learned feature exhibits robustness across different cross-modality datasets and matching settings.

Comparision with USVI-ReID Methods. The experimental results clearly demonstrate that our method outperforms existing unsupervised methods across various settings. Specifically, we achieve substantial improvements compared to single-modality unsupervised methods, with approximately 25% and 45% higher mAP scores on the SYSU-MM01 and RegDB tasks, respectively. Furthermore, when compared to H2H, which utilized an additional labeled RGB dataset for unsupervised cross-modality ReID, we achieve noticeable gains of 20.8% and 51.43% in mAP on SYSU-MM01 (all search) and RegDB (visible to infrared) tasks, respectively.

Ablation

The performance improvement of our method in USL-VI-ReID is primarily attributed to two key components: Modal-invariant Cluster Contrast learning (MICC) and Prototype Similarity Association (PSA) module. To assess the effectiveness of each component, we conducted ablation studies on the SYSU-MM01 and RegDB datasets. The results, presented in Table 1, validate the impact of these components.

Table 2: Ablation studies on the SYSU-MM01 dataset. "PSA" denotes the Prototype Similarity Association module. "MICC" means Modal-invariant Cluster Contrast learning. Rank at r accuracy(%), mAP(%) are reported.

Components			SYSU-MM01 (All Search)				
Baseline	PSA	MIC	$r1$	$r5$	$r10$	$r20$	mAP
√			40.2	69.8	82.0	91.1	37.6
√	√		51.2	79.1	89.3	95.1	48.1
√		√	49.9	78.8	88.5	94.3	46.7
√	√	√	53.3	81.1	90.8	96.5	50.2

Conclusion

This research paper introduces the task of unsupervised learning for visible-infrared re-identification (USL-VI-ReID), aiming to address the challenge of costly cross-modality annotations. To overcome the issue of significant cross-modality discrepancies in USL-VI-ReID, we propose an efficient collaborative unsupervised learning framework that leverages the concepts of homogenous joint learning and heterogeneous aggregation. This framework enhances unsupervised cross-modality recognition and has been extensively validated on two distinct tasks, demonstrating superior performance compared to current state-of-the-art unsupervised methods and even some supervised methods. These results pave the way for the practical deployment of unsupervised VI-ReID in real-world scenarios.

References

Chen, H.; Lagadec, B.; and Bremond, F. 2021. Ice: Inter-instance contrastive encoding for unsupervised person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14960–14969.

Chen, H.; Wang, Y.; Lagadec, B.; Dantcheva, A.; and Bremond, F. 2021. Joint Generative and Contrastive Learning for Unsupervised Person Re-identification. In *CVPR*.

Chen, Y.; Zhu, X.; and Gong, S. 2019. Instance-Guided Context Rendering for Cross-Domain Person Re-Identification. In *ICCV*.

Dai, P.; Ji, R.; Wang, H.; Wu, Q.; and Huang, Y. 2018. Cross-modality person re-identification with generative adversarial training. In *IJCAI*, volume 1, 6.

Dai, Z.; Wang, G.; Yuan, W.; Zhu, S.; and Tan, P. 2022a. Cluster contrast for unsupervised person re-identification. In *ACCV*, 1142–1160.

Dai, Z.; Wang, G.; Yuan, W.; Zhu, S.; and Tan, P. 2022b. Cluster contrast for unsupervised person re-identification. In *Proceedings of the Asian Conference on Computer Vision*, 1142–1160.

Dai, Z.; Wang, G.; Yuan, W.; Zhu, S.; and Tan, P. 2022c. Cluster contrast for unsupervised person re-identification. In *Proceedings of the Asian Conference on Computer Vision*, 1142–1160.

Ester, M.; Kriegel, H.; Sander, J.; and Xu, X. 1996a. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, Oregon, USA*, 226–231. AAAI Press.

Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X.; et al. 1996b. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, 226–231.

Ge, W.; Pan, C.; Wu, A.; Zheng, H.; and Zheng, W.-S. 2021. Cross-camera feature prediction for intra-camera supervised person re-identification across distant scenes. In *ACMMM*, 3644–3653.

Guo, J.; Yuan, Y.; Huang, L.; Zhang, C.; Yao, J.-G.; and Han, K. 2019. Beyond human parts: Dual part-aligned representations for person re-identification. In *ICCV*, 3642–3651.

Hao, X.; Zhao, S.; Ye, M.; and Shen, J. 2021. Cross-modality person re-identification via modality confusion and center aggregation. In *CVPR*, 16403–16412.

Liang, W.; Wang, G.; Lai, J.; and Xie, X. 2021. Homogeneous-to-Heterogeneous: Unsupervised Learning for RGB-Infrared Person Re-Identification. *IEEE Transactions on Image Processing*.

Mang Ye, J. L., Xiangyuan Lan; and Yuen, P. 2018. Hierarchical discrimi-native learning for visible thermal person re-identification. In *AAAI*, 7501–7508.

Park, D. T. N. H. G. H. . R. 2017. Person Recognition System Based on a Combination of Body Images from Visible Light and Thermal Cameras. *Sensors*, 17: 605.

Radenovic, F.; Tolias, G.; and Chum, O. 2019. Fine-Tuning CNN Image Retrieval with No Human Annotation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(7): 1655–1668.

Tian, X.; Zhang, Z.; Lin, S.; Qu, Y.; Xie, Y.; and Ma, L. 2021. Farewell to mutual information: Variational distillation for cross-modal person re-identification. In *CVPR*, 1522–1531.

Wang, J.; Zhang, Z.; Chen, M.; Zhang, Y.; Wang, C.; Sheng, B.; Qu, Y.; and Xie, Y. 2022. Optimal Transport for Label-Efficient Visible-Infrared Person Re-Identification. In *ECCV*, 93–109.

Wang, M.; Lai, B.; Huang, J.; Gong, X.; and Hua, X.-S. 2021. Camera-aware proxies for unsupervised person re-identification. In *AAAI*, 2764–2772.

Wang, Z.; Hu, R.; Chen, C.; Yu, Y.; Jiang, J.; Liang, C.; and Satoh, S. 2017. Person reidentification via discrepancy matrix and matrix metric. *IEEE transactions on cybernetics*, 48(10): 3006–3020.

Wang, Z.; Wang, Z.; Zheng, Y.; Chuang, Y.-Y.; and Satoh, S. 2019a. Learning to Reduce Dual-Level Discrepancy for Infrared-Visible Person Re-Identification. In *CVPR*.

Wang, Z.; Wang, Z.; Zheng, Y.; Wu, Y.; Zeng, W.; and Satoh, S. 2019b. Beyond intra-modality: A survey of heterogeneous person re-identification. *arXiv preprint arXiv:1905.10048*.

Wei, L.; Zhang, S.; Gao, W.; and Tian, Q. 2018. Person Transfer GAN to Bridge Domain Gap for Person Re-Identification. In *CVPR*.

Wu, A.; Zheng, W.; Yu, H.; Gong, S.; and Lai, J. 2017a. RGB-Infrared Cross-Modality Person Re-identification. In *ICCV*, 5390–5399.

Wu, A.; Zheng, W.-S.; Yu, H.-X.; Gong, S.; and Lai, J. 2017b. RGB-infrared cross-modality person re-identification. In *ICCV*, 5380–5389.

Wu, A.; Zheng, W.-S.; Yu, H.-X.; Gong, S.; and Lai, J. 2017c. RGB-Infrared Cross-Modality Person Re-identification. In *ICCV*.

Wu, Q.; Dai, P.; Chen, J.; Lin, C.-W.; Wu, Y.; Huang, F.; Zhong, B.; and Ji, R. 2021. Discover Cross-Modality Nuances for Visible-Infrared Person Re-Identification. In *CVPR*.

Ye, M.; Shen, J.; J. Crandall, D.; Shao, L.; and Luo, J. 2020. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In *ECCV*, 229–247. Springer.

Ye, M.; Shen, J.; Lin, G.; Xiang, T.; Shao, L.; and Hoi, S. C. H. 2022. Deep Learning for Person Re-Identification: A Survey and Outlook. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(6): 2872–2893.

Zhang, X.; Ge, Y.; Qiao, Y.; and Li, H. 2021. Refining Pseudo Labels with Clustering Consensus over Generations for Unsupervised Object Re-identification. In *CVPR*.

Zhang, X.; Li, D.; Wang, Z.; Wang, J.; Ding, E.; Shi, J.; Zhang, Z.; and Wang, J. 2022. Implicit Sample Extension for Unsupervised Person Re-Identification.