Cross-Modality Person Re-identification with Memory-Based Contrastive Embedding

De Cheng[1*], Xiaolong Wang[1*], Nannan Wang[1†], Zhen Wang[2], Xiaoyu Wang[3], Xinbo Gao[4]

[1] Xidian University,
[2] Zhejiang Lab,
[3] University of Science and Technology of China,
[4] Chongqing University of Posts and Telecommunications
{dcheng,nnwang}@xidian.edu.cn

Abstract

Visible-infrared person re-identification (VI-ReID) aims to retrieve the person images of the same identity from the RGB to infrared image space, which is very important for real-world surveillance system. In practice, VI-ReID is more challenging due to the heterogeneous modality discrepancy, which further aggravates the challenges of traditional single-modality person ReID problem, i.e., inter-class confusion and intra-class variations. In this paper, we propose an aggregated memory-based cross-modality deep metric learning framework, which benefits from the increasing number of learned modality-aware and modality-agnostic centroid proxies for cluster contrast and mutual information learning. Furthermore, to suppress the modality discrepancy, the proposed cross-modality alignment objective simultaneously utilizes both historical and up-to-date learned cluster proxies for enhanced cross-modality association. Such training mechanism helps to obtain hard positive references through increased diversity of learned cluster proxies, and finally achieves stronger "pulling close" effect between cross-modality image features. Extensive experiment results demonstrate the effectiveness of the proposed method, surpassing state-of-the-art works significantly by a large margin on the commonly used VI-ReID datasets.

Introduction

Person re-identification (ReID) aims at retrieving a person of interest from a large-scale image gallery set, captured across multiple non-overlapping cameras (Cheng et al. 2016). It plays an important role in video surveillance, security, and pedestrian analysis, and has obtained great attention over the past decades. Conventional person Re-ID mainly focuses on single-modality, i.e., all person images are RGB images taken by visible cameras during day time. In recent years, impressive performances have been obtained on most benchmark datasets (Fu et al. 2021; Cheng et al. 2022), even under the unsupervised learning scenarios. However, the visible cameras cannot image clearly under poor lighting conditions, e.g., in the dark environment, which greatly limits the application of person Re-ID in the real-world surveillance system. To overcome this obstacle, many infrared (IR) images are equipped in surveillance scenarios to overcome the illumination variants under different lighting conditions, which assists the visible cameras. Therefore, this greatly shows the importance of VI-ReID for real-world surveillance system.

In practice, VI-ReID is more challenging due to the heterogeneous modality discrepancy, which is caused by different wavelength of RGB and IR images captured with different imaging equipments. Such modality discrepancy further aggravates the challenges of traditional person ReID problem, i.e., inter-class confusion and intra-class variations (e.g., pose, viewpoints, illumination, background cluster, occlusion, etc.). Some detailed analysis (Liu et al. 2022) points out that, sometimes the image features corresponding to different identifies under the same modality are even similar than those with the same identity but under different modalities. Therefore, the key solution to cross-modality object recognition is to achieve modality unification, either from the image-level or feature-level (Wei et al. 2021), or their combinations. Among all these methods, we have found that the feature-level modality-alignment-based methods (or their combinations) show more impressive experiment results for the cross-modality object recognition task. As for the feature-level alignment, existing studies mainly focus on learning modality shared or invariant features. The representative works include the dual-path networks (Lu et al. 2020) to learn modality-specific and modality-shared feature representations, and the one-stream weight-sharing network (Hao et al. 2019) to directly extract modality-shared features.

To reduce the modality discrepancy in feature level, we propose a aggregated memory-based deep metric learning framework for cross-modality VI-ReID. Recently, memory-based contrastive learning has been extensively explored in deep metric learning(Dai et al. 2021; He et al. 2020; Deng et al. 2021; Liu et al. 2022). We analyze that most of these successes owe to the increase of negative features from the memory bank, which greatly assists the contrastive learning. While in our work, the memory-based cross-modality deep metric learning not only benefits from the increasing modality-aware and modality-agnostic negative examples for cluster contrast, but also the model drift phenomena (Wang et al. 2020; Liu et al. 2022), which adopts the historical learned modality-aware and modality-agnostic proxies to enhance modality alignment with hard positive references, and

[*]Equal contribution.
[†]Corresponding author.

consequentially suppress the modality discrepancy.

Specifically, the proposed memory-based cross-modality deep metric learning consists of three components: the modality-aware and modality-agnostic cluster contrast objective, the historical memory centroid based cross-modality mutual information constraint, as well as the global modality-agnostic feature matching. In the first component, the proposed method learns three modality-aware proxies (RGB, IR and auxiliary modality) and one modality-agnostic proxy for each identity, simultaneously. To further enhance the cross-modality association, we propose the cross-modality mutual information constraint and the global modality-agnostic feature matching objective, assisted by the model-drift proxies (i.e., historical modality-aware and modality-agnostic proxies). Obviously, we can clearly see that our proposed cross-modality deep metric learning simultaneously utilizes both historical and up-to-date learned proxies for cluster contrast based modality alignment, which increases the diversity of proxies in the memory bank. Besides, since these historical cluster centroids are relatively farther away from the modality/identity boundary than up-to-date dynamically learned proxies stored in the memory bank, they could help to get hard positive references and result in stronger "pulling close" effect between cross-modality image features with the same identity. Thus, the modality discrepancy can be greatly suppressed, and we can learn modality-shared features while keeping high discriminability.

The contributions can be summarized as follows:

- This paper proposes an aggregated memory-based cross-modality deep metric learning framework, which benefits from the increasing number of modality-aware and modality-agnostic proxies for cluster contrast and mutual information learning.
- To suppress the modality discrepancy, the proposed cross-modality alignment objective simultaneously utilizes both historical and up-to-date learned proxies for enhanced cross-modality association. This mechanism helps to get hard positive references through increased diversity of learned proxies, and finally achieves stronger "pulling close" effect between cross-modality image features.
- Extensive experiment results demonstrate the effectiveness of the proposed method, which surpass the state-of-the-art methods significantly by a large margin on the commonly used VI-ReID datasets.

Related Work

Visible-Infrared Person Re-identification (VI-ReID) aims to retrieve person images with the same identity from RGB to infrared image space, or vice versa. The main challenge is the heterogeneous modality discrepancy. Therefore, almost all works devote to achieving modality alignment through suppressing the modality discrepancy and learning modality-shared feature representations. Here, we roughly divide existing works into the following categories: **feature-level** and **image-level** modality alignment algorithms.

Feature-Level Modality Alignment aims to learn modality-shared feature representations, which mitigates the modality difference in feature level. The following lists some representative works. (Wu et al. 2017) proposed a deep zero-padding network to align data of different modalities. (Ye et al. 2018) adopts a two-stream network architecture to extract features of different modalities, and uses a combination of feature learning and metric learning to compensate for the differences in modalities. (Dai et al. 2018) proposed a cross-modal generative adversarial network (GAN), which utilizes the generator to learn features under different modalities, and uses the discriminator to classify the modalities. (Ye et al. 2020) considers both intra-modality feature connections and inter-modality adjacent structural information, for dual-attentive aggregation learning. (Wu et al. 2021) proposed a modality alignment network to discover cross-modality nuances for VI-ReID. Overall, these methods mainly focus on suppressing modality discrepancy though aligning the distributions of cross-modality features.

Image-Level Modality Alignment approaches try to alleviate modality discrepancy by generating some intermediate modality images. The representative GAN-based methods for modality transfer include AlignGAN (Wang et al. 2019a), D2RL (Wang et al. 2019b). AlignGAN proposed a pixel and feature alignment network to generate IR image from RGB image. D2RL introduced a dual-level discrepancy reduction strategy by training an image-level sub-network to translate a RGB image to its infrared counterpart, and vice versa. There also contain some works trying to generate intermediate modality images for joint learning to achieve modality alignment, such as X-modality (Li et al. 2020), CAJ (Ye et al. 2021), SMCL (Wei et al. 2021). Specifically, X-modal adopts a lightweight generator to generate intermediate modality from RGB images. CAJ randomly extracts a channel from RGB images and expands it to three dimensions as auxiliary modality. SMCL applies RGB and IR images to generate an intermediate modality and learn it jointly to guide the generation of modality-invariant representations. Usually, GAN-based methods are often difficult to train and the network is more complex, while the methods based on intermediate modality are often simpler and more effective.

There also contain some **other methods for VI-ReID.** (Chen et al. 2021) proposed a neural feature search strategy to select features that can reduce modality difference while keeping feature discriminability. (Fu et al. 2021) proposed a Batch-Norm oriented cross-modality neural architecture search method. Such methods are theoretically feasible, but are often difficult to get optimal solution in practice.

Memory-Based Deep Metric Learning has been extensively explored in supervised, semi-supervised and unsupervised learning recently (Dai et al. 2021; He et al. 2020; Deng et al. 2021; Liu et al. 2022). Contrastive learning based on memory bank has achieved good results in many scenarios(Dai et al. 2021; He et al. 2020). In MOCO(He et al. 2020), it builds a large and consistent dictionary on-the-fly to facilitate contrastive unsupervised learning. The MUAM (Liu et al. 2022) method proposed to learn memory-augmented unidirectional metrics for cross-modality person ReID. In (Deng et al. 2021), it proposed a variational prototype learning mechanism with memorized feature being injected into the prototypes, for face recognition. Different from existing

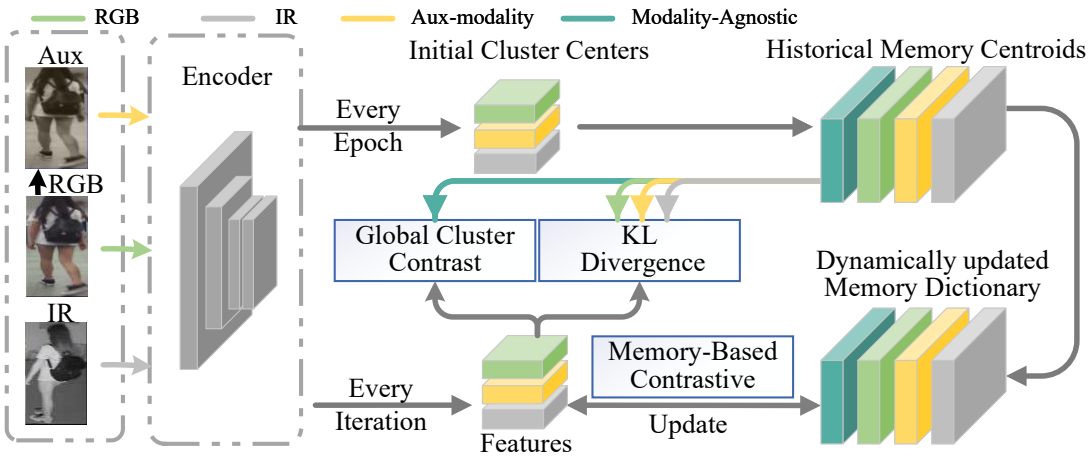

Figure 1: Overall framework of the proposed aggregated memory-based deep metric learning for VI-ReID. The framework contains the weight-sharing backbone network, the historical and dynamically updated memory bank to store the modality-aware and modality-agnostic centroids. The network is optimized simultaneously by the memory-based cluster contrast, cross-modality mutual information constraint and the global modality-agnostic cluster contrast objective.

memory-bank based supervised contrastive learning, our proposed cross-modality deep metric learning simultaneously utilizes both historical and up-to-date learned proxies for cluster-contrast-based modality alignment, which increases the diversity of proxies in the memory bank.

The Proposed Method

Problem Definition

Let $\mathcal{X}^v = \{\mathbf{x}_i^v\}_{i=1}^{N^v}$ and $\mathcal{X}^r = \{\mathbf{x}_i^r\}_{i=1}^{N^r}$ denote the respective visible RGB images and the infrared images in a cross modality dataset, where N^v and N^r are the number of RGB and IR images, then the total number of training images is $N = N^v + N^r$. The corresponding ground-truth label space can be denoted as $\mathcal{Y} = \{\mathbf{y}_i\}_{i=1}^{N^p}$, where N^p is the total number of person identities in the dataset. Given a certain query person image in one modality, cross-modality person ReID task aims to retrieve the pedestrian images in another modality with the same identity, according to the learned image feature similarities.

As shown in Figure 1, it illustrates the overall framework of the proposed aggregated memory-based deep matric learning for IR-ReID. The framework contains the weight-sharing backbone network, the historical and dynamically updated memory bank to store the modality-aware and modality-agnostic centroids. Specifically, we adopt Resnet50 (He et al. 2016) pre-trained on ImageNet dataset to work as our backbone network. Then, each image corresponds to one feature vector. Here, we denote $\mathbf{f}_i^v \in \mathbb{R}^{2048\times 1}$ and $\mathbf{f}_i^r \in \mathbb{R}^{2048\times 1}$ as the i-th extracted features for the corresponding images $\mathbf{x}_i^v$ and $\mathbf{x}_i^r$, respectively. The CNN architecture is jointly optimized by the memory-based modality-aware and modality-agnostic cluster contrast objective, the cross-modality mutual information constraint, and the global modality-agnostic cluster contrast objective.

We note that the intermediate modality is very useful for reducing the modality discrepancy in VI-ReID. In our work, we also adopt the channel augmentation strategy proposed by (Ye et al. 2021) as the auxiliary modality. Therefore, our framework contains three modality inputs, i.e., RGB, IR, and the auxiliary modality which is generated by randomly exchanging the color channels of the RGB image. We denote $\mathcal{X}^a = \{\mathbf{x}_i^a\}_{i=1}^{N^a}$ as the auxiliary modality dataset corresponding to the visible modality, where $N^a = N^v$, and its corresponding feature vector is denoted as $\mathbf{f}_i^a \in \mathbb{R}^{2048\times 1}$.

Learning Modality-Aware and Modality-Agnostic Proxies

In the vallina memory-based cluster contrast learning (Dai et al. 2021; Yao and Xu 2021), the most important component is the cluster-based memory bank, where each cluster is represented by a mean feature vector $\mathcal{W}^m$ (also denoted as cluster center), and all the cluster feature vectors are updated based on the individual features. In the following, we take the training images in the RGB modality as example. Given the training images $\mathcal{X}^v = \{\mathbf{x}_i^v\}_{i=1}^{N^v}$, we can extract the corresponding features $\mathcal{F}^v = \{\mathbf{f}_i^v\}_{i=1}^{N^v}$. Then, the cluster memory bank $\mathcal{W}^v \in \mathbb{R}^{d\times N^p}$ for the RGB modality is initialized by the mean of feature vectors in each cluster, where d and N^p represent the feature dimension and number of identities/-clusters, respectively. Specifically, the cluster centroid $\mathcal{W}_k^v$ for the k-th class can be initialized as follows,

$$\mathcal{W}_k^v = \frac{1}{|\mathcal{F}_k^v|} \sum_{\mathbf{f}_i^v \in \mathcal{F}_k^v} \mathbf{f}_i^v, \tag{1}$$

where $\mathcal{F}_k^v$ denotes the subset of training images belonging to the k-th class in the RGB feature space, $|\cdot|$ represents the number of instances in the cluster set. The cluster centroids stored in the memory bank are updated with the corresponding cluster feature vectors during model training.

Then, the memory-bank-based cluster-wise contrastive learning can be derived as Eq. 2, which also acts as a non-

parametric classifier.

$$\mathcal{L}^v = -\log \frac{\exp(\mathbf{f}_i^v \cdot \mathcal{W}_{y_i}^v/\tau)}{\sum_{k=1}^{N^p} \exp(\mathbf{f}_i^v \cdot \mathcal{W}_k^v/\tau)}, \tag{2}$$

where τ is the temperature hyper-parameter, y_i is the corresponding ground-truth label for the image feature $\mathbf{f}_i^v$, and $\mathcal{W}_{y_i}^v$ is its positive cluster centroid stored in current memory bank. The objective function $\mathcal{L}^v$ encourages the feature $\mathbf{f}_i^v$ to have higher similarity with its corresponding ground-truth cluster centroid $\mathcal{W}_{y_i}^v$ and dissimilarity with the other $N^p - 1$ cluster centroids.

During model training process, we adopt the mean of instance features belonging to one specific cluster in a mini-batch to momentum update the corresponding cluster center in the memory bank, as follows,

$$\mathcal{W}_{y_i}^v \leftarrow \gamma \mathcal{W}_{y_i}^v + (1-\gamma)\overline{\mathbf{f}}_{y_i}^v, \tag{3}$$

where γ is the momentum updating factor, $\mathcal{W}_{y_i}^v$ is the y_i-th cluster centroid in the memory bank, $\overline{\mathbf{f}}_{y_i}^v$ is the mean of instance features belonging to the y_i-th class in current mini-batch.

Optimizing the memory-based cluster contrast objective function $\mathcal{L}^v$ in the RGB modality, we can obtain the corresponding parameters of the centroids in the memory bank $\mathcal{W}^v = \{\mathcal{W}_k^v\}_{k=1}^{N^p}$. Corresponding, we can optimize the whole network architecture with other modality data in the same way, simultaneously. The loss function and centroid parameter set can be represented as $\mathcal{L}^r$, $\mathcal{W}^r = \{\mathcal{W}_k^r\}_{k=1}^{N^p}$ and $\mathcal{L}^a$, $\mathcal{W}^a = \{\mathcal{W}_k^a\}_{k=1}^{N^p}$, for the IR modality and auxiliary modality, respectively. Meanwhile, we also construct a modality-agnostic brunch to suppress the modality discrepancy with all the training data in the same way, and the corresponding loss function and parameter set can be represented as $\mathcal{L}^u$, $\mathcal{W}^u = \{\mathcal{W}_k^u\}_{k=1}^{N^p}$.

Therefore, the memory-based cluster contrast learning under the whole cross-modality data arrives at,

$$\mathcal{L}_{\mathcal{W}} = \mathcal{L}^v + \mathcal{L}^r + \mathcal{L}^a + \mathcal{L}^u. \tag{4}$$

Cross-Modality Mutual Information

To further enhance the modality association and suppress the modality discrepancy, we also propose the cross-modality mutual information constraint. As previously introduced, we have learned the identity centroids for each modality separately. That is to say, the centroids stored in the memory bank only learn knowledge from their corresponding modalities. Therefore, given a pedestrian image feature (denoted as $\mathbf{f}$), no matter which modality it belongs to, if its relatively nearest centroids in different modalities correspond to the same identity, it means that our model extracts modality-shared image features and the modality discrepancy is eliminated.

Specifically, we propose the cross-modality mutual information constraint according to the Kullback-Leibler divergence (Wu et al. 2021). To this end, we first transform the obtained image feature $\mathbf{f}_i^v$ into the probability response, based on the learned image centroids in different modalities.

In the memory-based cluster contrast learning, we have observed that we can benefit more from the historical learned cluster centroids. Because some historical cluster centroids are relatively farther away from the modality/identity boundary than up-to-date dynamically learned centroids in the memory bank, which could result in enhanced "pulling close" effect on the counterpart-modality distributions. Therefore, we propose to use the centroids learned at the end of previous epoch to act as the classifier, thus the parameters of the centroids in the following training epoch always keep fixed.

To be specific, the probability response $P_v^v\left(\mathbf{f}_i^v \mid \mathbf{C}^v\right)$ for the input feature $\mathbf{f}_i^v$ under the cluster centroids in the RGB modality $\mathbf{C}^v = \{\mathbf{c}_k^v\}_{k=1}^{N^p}$, can be denoted as follows,

$$P_v^v\left(\mathbf{f}_i^v \mid \mathbf{C}^v\right) = \frac{\exp\left(\mathbf{f}_i^v \cdot \mathbf{c}_+^v/\tau\right)}{\sum_{k=1}^{N^p} \exp\left(\mathbf{f}_i^v \cdot \mathbf{c}_k^v/\tau\right)}. \tag{5}$$

Correspondingly, the probability response $P_r^v\left(\mathbf{f}_i^v \mid \mathbf{C}^r\right)$ for the input feature $\mathbf{f}_i^v$ under the cluster centroids in the IR modality $\mathbf{C}^r = \{\mathbf{c}_k^r\}_{k=1}^{N^p}$, can be denoted as follows,

$$P_r^v\left(\mathbf{f}_i^v \mid \mathbf{C}^r\right) = \frac{\exp\left(\mathbf{f}_i^v \cdot \mathbf{c}_+^r/\tau\right)}{\sum_{k=1}^{N^p} \exp\left(\mathbf{f}_i^v \cdot \mathbf{c}_k^r/\tau\right)}. \tag{6}$$

Please note that, the historical cluster centroid parameters $\mathbf{C} = \{\mathbf{C}^v, \mathbf{C}^r, \mathbf{C}^a, \mathbf{C}^u\}$ here are a little different from the centroids stored in up-to-date memory bank $\mathcal{W} = \{\mathcal{W}^v, \mathcal{W}^r, \mathcal{W}^a, \mathcal{W}^u\}$, though both of which could represent the cluster centers of the identities. The historical cluster centroids $\mathbf{C} = \{\mathbf{C}^v, \mathbf{C}^r, \mathbf{C}^a, \mathbf{C}^u\}$ are computed by the mean of feature vectors within the same identity in its own modality space, where the feature vectors are extracted by the model obtained at the end of the previous training epoch.

Therefore, the Kullback-Leibler divergence (Wu et al. 2021) between probability responses in the RGB modality and the IR modality arrives at,

$$\begin{aligned}\mathcal{L}_{MI}^{v\leftrightarrow r} = P_r^v\left(\mathbf{f}_i^v \mid \mathbf{C}^r\right) \log \frac{P_r^v\left(\mathbf{f}_i^v \mid \mathbf{C}^r\right)}{P_v^v\left(\mathbf{f}_i^v \mid \mathbf{C}^v\right)} + \\ P_v^r\left(\mathbf{f}_i^r \mid \mathbf{C}^v\right) \log \frac{P_v^r\left(\mathbf{f}_i^r \mid \mathbf{C}^v\right)}{P_r^r\left(\mathbf{f}_i^r \mid \mathbf{C}^r\right)}.\end{aligned} \tag{7}$$

Obviously, objective function in Eq. 7, encourages the input feature vector $\mathbf{f}_i^v$ in the RGB modality to have the consistency probability response on both its original RGB modality centroids and the IR modality centroids, and vice versa for the input features in another IR modality. By this way, we can encourage the model to learn knowledge among different modalities, and further learn modality-irrelevant features.

As our framework contains another auxiliary modality data $\mathcal{X}^a$, which is generated from the corresponding RGB modality data $\mathcal{X}^v$ through the channel augmentation strategy, we also build the mutual information constraint between the auxiliary modality and the IR modality to further suppress the modality discrepancy. It can be denoted as $\mathcal{L}_{MI}^{a\leftrightarrow r}$, which is built in the same way as $\mathcal{L}_{MI}^{v\leftrightarrow r}$. Therefore, the overall cross-modality mutual information constraint can be written as,

$$\mathcal{L}_{MI} = \mathcal{L}_{MI}^{v\leftrightarrow r} + \mathcal{L}_{MI}^{a\leftrightarrow r}. \tag{8}$$

Global Modality-Agnostic Cluster Contrast

To further suppress the modality discrepancy, we also propose the global modality-agnostic cluster contrast learning, on top of the learned historical modality-agnostic cluster centroids $\mathbf{C}^u$. The objective function can be expressed as follows,

$$\mathcal{L}_{GC} = \max\left[\left\|\mathbf{f}_i - \mathbf{c}^u_{y_i}\right\|_2 - \min_{k\neq y_i}\|\mathbf{f}_i - \mathbf{c}^u_k\|_2 + \alpha, 0\right], \tag{9}$$

where $\mathbf{f}_i$ can be any input feature vector in all the modality space, y_i is its corresponding label/identity, $\mathbf{C}^u = \{\mathbf{c}^u_k\}^{N^p}_{k=1}$ is the learned historical modality agnostic cluster centroids. $\mathbf{c}^u_{y_i}$ is the positive centroid corresponding to feature $\mathbf{f}_i$, and α is the least margin parameter between the positive distance and the minimum negative distance. Through minimizing the above loss function, we can suppress the modality discrepancy in an overall modality-agnostic way.

Overall Objective Function

Finally, the proposed method will consider both the modality-aware and modality-agnostic, as well as the historical and up-to-date memory-based cluster contrast learning, for efficient feature-level modality alignment. The overall objective function $\mathcal{L}$ can be expressed as follows,

$$\mathcal{L} = \mathcal{L}_{\mathcal{W}} + \lambda_1\mathcal{L}_{MI} + \lambda_2\mathcal{L}_{GC}, \tag{10}$$

where λ_1 and λ_2 are two hyper-parameters to balance the above three terms in the overall loss function.

Experiment

Dataset and Evaluation Protocol

Dataset. We evaluate our proposed method on two widely used VI-ReID datasets: SYSU-MM01 (Wu et al. 2017) and RegDB (Nguyen et al. 2017).

• **SYSU-MM01** is one relatively large benchmark VI-ReID dataset, including 491 identities collected from four visible and two near-infrared cameras. The training set contain 395 identities with 22,258 RGB and 11,909 infrared images, and the test set contain the remaining 96 identities. Following (Wu et al. 2017), we evaluate the proposed method under two search modes, the All-search mode and the Indoor-search mode, every mode contains single-shot and multi-shot settings. In All-search mode, the gallery set contains all visible images, while for indoor-search mode, the gallery set only contain images captured from indoor cameras. In both of these two modes, the query set contain all infrared testing images. For single-shot setting, we randomly sample 301 images of 96 identities in the gallery set, while 3010 images in multi-shot setting.

• **RegDB** dataset is captured by a pair of aligned visible and infrared cameras (Nguyen et al. 2017). It contains 412 identities with 8240 images, where randomly selecting 206 identities for training and the remaining 206 identities for testing, and each identity corresponds to 10 RGB and 10 infrared images. During model evaluation, we adopt the commonly used two experiment settings on RegDB dataset: Visible-to-Infrared and Infrared-to-Visible, representing querying visible image from infrared image gallery, and vice versa.

Evaluation Protocol. All experiments follow the common evaluation protocols used for VI-ReID (Wu et al. 2017; Ye et al. 2018). The Cumulative Matching Characteristic (CMC) and mean Average Precision (mAP) are adopted as evaluation matrix. We run the code five times under each experiment setting, and report the mean precision for all experiments.

Implementation Details. We adopt ResNet-50 (He et al. 2016) pre-trained on ImageNet as our backbone network. Our model is implemented by PyTorch and trained on a single RTX3090 GPU platform. Following CAJ (Ye et al. 2021), We inserted nonlocal (Wang et al. 2018) structure inside Resnet-50 network architecture. During model optimization, we adopt the Adam optimization method, the weight decay is set to 0.0005, the initial learning rate is set to 0.00035 with a warmup strategy, and it is divided by 10 at the 20-th and 40-th epochs. The input images are re-scaled to the size of 384×128 with data augmentation like randomly flipping and erasing used. We train the whole model for 80 epochs overall. For each training step, the batch-size is set to 64, where we randomly sample 8 identities, and each with 4 RGB and 4 infrared images. The momentum update factor in Eq. 3 is set to 0.1 for the modality-agnostic centers and 0.3 for other three modality-aware centers. The hyper-parameters λ_1 and λ_2 in Eq. 10 is set to 1.2 and 1.0, respectively. All the temperature parameter τ in Eq. 2 5 6 is set to 0.05, and α in Eq. 9 is set to 0.3, empirically.

Comparison with State-of-the-Art Methods

We compare the proposed method with existing state-of-the-art approaches, including Zero-Pad(Wu et al. 2017), HSME(Hao et al. 2019), D2RL(Wang et al. 2019b), Align-GAN(Wang et al. 2019a), X-modal(Li et al. 2020), cm-SSFT(Lu et al. 2020), DDAG(Ye et al. 2020), HAT(Ye, Shen, and Shao 2020), CM-NAS(Fu et al. 2021), MPANet(Wu et al. 2021), CAJ(Ye et al. 2021), FMCNet(Zhang et al. 2022), MAUM(Liu et al. 2022).

Results on SYSU-MM01 Dataset. As shown in Table 1, the proposed method yields the best results, outperforming all existing state-of-the-art methods. In the four experiment settings: All-search Single-shot, All-search Multi-shot, Indoor-search Single-shot, Indoor-search Multi-shot, our method achieves 72.01%, 65.77%, 86.06% and 80.64% in terms of mAP, surpassing the second best method by a margin of 3.22%, 2.86%, 4.12% and 5.53%, respectively.

Results on RegDB Dataset. As shown in Table 2, our proposed method also achieves the best performances over all the compared methods. In the Visible-to-Infrared experiment setting, we achieve 93.15% and 88.32% in terms of Rank-1 and mAP evaluation matrices, surpassing the second best by a margin of 4.03% and 3.23%, respectively. While for the Infrared-to-Visible experiment setting, we achieve 93.42% Rank-1 accuracy and 87.95% mAP, surpassing the second best by a margin of 5.04% and 3.61%, respectively.

Ablation Study

The proposed cross-modality deep metric learning framework contains three novel ingredients corresponding to three terms as shown in Eq. 10: 1) the modality-aware and modality-agnostic cluster contrast objective $\mathcal{L}_{\mathcal{W}}$; 2) the historical

Methods	Venue	All-search				Indoor-search			
		Single-shot		Multi-shot		Single-shot		Multi-shot	
		Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
Zero-Pad(Wu et al. 2017)	ICCV-17	14.80	15.95	19.13	10.89	20.58	26.92	24.43	18.86
HSME(Hao et al. 2019)	AAAI-19	20.68	23.12	-	-	-	-	-	-
D2RL(Wang et al. 2019b)	CVPR-19	28.9	29.2	-	-	-	-	-	-
AlignGAN(Wang et al. 2019a)	ICCV-19	42.40	40.70	51.50	33.90	45.90	54.30	57.10	45.30
X-Modal(Li et al. 2020)	AAAI-20	49.92	50.73	-	-	-	-	-	-
cm-SSFT(Lu et al. 2020)	CVPR-20	61.60	63.20	63.40	62.00	70.50	72.60	73.00	72.40
DDAG(Ye et al. 2020)	ECCV-20	54.75	53.02	-	-	61.02	67.98	-	-
HAT(Ye, Shen, and Shao 2020)	TIFS-20	55.29	53.89	-	-	62.10	69.37	-	-
CM-NAS(Fu et al. 2021)	CVPR-21	61.99	60.02	68.68	53.45	67.01	72.95	76.48	65.11
MPANet(Wu et al. 2021)	CVPR-21	70.58	68.24	<u>75.58</u>	<u>62.91</u>	76.74	80.95	<u>84.22</u>	<u>75.11</u>
CAJ(Ye et al. 2021)	ICCV-21	69.88	66.89	-	-	76.26	80.37	-	-
FMCNet(Zhang et al. 2022)	CVPR-22	66.34	62.51	73.44	56.06	68.15	74.09	78.86	63.82
MAUM(Liu et al. 2022)	CVPR-22	<u>71.68</u>	<u>68.79</u>	-	-	<u>76.97</u>	<u>81.94</u>	-	-
Ours	-	**74.77**	**72.01**	**78.35**	**65.77**	**83.48**	**86.06**	**88.43**	**80.64**

Table 1: Comparison with the state-of-the-art VI-ReID methods on SYSU-MM01 dataset.

Methods	Venue	Visible-to-Infrared				Infrared-to-Visible			
		Rank-1	Rank-10	Rank-20	mAP	Rank-1	Rank-10	Rank-20	mAP
Zero-Pad(Wu et al. 2017)	ICCV-17	17.75	34.21	44.35	18.90	16.63	34.68	44.25	17.82
HSME(Hao et al. 2019)	AAAI-19	50.85	73.36	81.66	47.00	50.15	72.40	81.07	46.16
D2RL(Wang et al. 2019b)	CVPR-19	43.4	66.1	76.3	44.1	-	-	-	-
AlignGAN(Wang et al. 2019a)	ICCV-19	57.9	-	-	53.6	56.3	-	-	53.4
X-Modal(Li et al. 2020)	AAAI-20	62.21	83.13	91.72	60.18	-	-	-	-
cm-SSFT(Lu et al. 2020)	CVPR-20	72.3	-	-	72.9	71.0	-	-	71.7
DDAG(Ye et al. 2020)	ECCV-20	69.34	86.19	91.49	63.46	68.06	85.15	90.31	61.80
HAT(Ye, Shen, and Shao 2020)	TIFS-20	71.83	87.16	92.16	67.56	70.02	86.45	91.61	66.30
CM-NAS(Fu et al. 2021)	CVPR-21	84.54	95.18	<u>97.85</u>	80.32	82.57	94.51	97.37	78.31
MPANet(Wu et al. 2021)	CVPR-21	83.7	-	-	80.9	82.8	-	-	80.7
CAJ(Ye et al. 2021)	ICCV-21	85.03	<u>95.49</u>	97.54	79.14	84.75	<u>95.33</u>	<u>97.51</u>	77.82
FMCNet(Zhang et al. 2022)	CVPR-22	<u>89.12</u>	-	-	84.43	<u>88.38</u>	-	-	83.86
MAUM(Liu et al. 2022)	CVPR-22	87.87	-	-	<u>85.09</u>	86.95	-	-	<u>84.34</u>
Ours	-	**93.15**	**96.91**	**98.54**	**88.32**	**93.42**	**97.10**	**98.53**	**87.95**

Table 2: Comparison with state-of-the-art VI-ReID methods on RegDB dataset.

Methods	Channels	SYSU-MM01(Single-shot)								RegDB(Visible-to-Infrared)			
		All-search				Indoor-search							
		Rank-1	Rank-10	Rank-20	mAP	Rank-1	Rank-10	Rank-20	mAP	Rank-1	Rank-10	Rank-20	mAP
$\mathcal{L}_{\mathcal{W}}$	2	57.81	92.65	97.14	57.45	62.30	95.76	98.91	69.73	89.68	95.53	97.77	85.27
$\mathcal{L}_{\mathcal{W}} + \mathcal{L}_{MI}$	2	69.79	95.48	98.64	67.65	79.99	98.51	99.46	83.09	92.01	96.59	98.36	87.15
$\mathcal{L}_{\mathcal{W}} + \mathcal{L}_{MI} + \mathcal{L}_{\mathcal{GC}}$	2	69.94	95.35	98.70	67.92	79.94	98.71	99.65	83.23	91.78	96.56	98.38	87.45
CE	3	65.97	94.17	97.79	62.61	71.53	97.25	99.47	76.62	83.20	93.69	96.26	77.18
$\mathcal{L}^u$	3	67.93	95.20	98.45	65.58	76.61	98.52	99.76	80.94	91.59	96.34	97.99	86.49
$\mathcal{L}_{\mathcal{W}}$	3	66.99	94.44	98.09	64.85	75.74	97.61	99.32	79.82	91.16	96.28	98.05	87.16
$\mathcal{L}_{\mathcal{W}} + \mathcal{L}^D_{MI}$	3	71.95	96.35	98.78	68.80	79.81	98.79	99.84	83.11	93.03	96.74	98.30	88.20
$\mathcal{L}_{\mathcal{W}} + \mathcal{L}_{MI}$	3	73.97	96.64	99.13	71.27	83.68	98.52	99.54	85.69	93.05	96.96	98.47	88.28
$\mathcal{L}_{\mathcal{W}} + \mathcal{L}^D_{MI} + \mathcal{L}^D_{\mathcal{GC}}$	3	71.88	96.46	98.72	69.12	80.10	98.66	99.81	83.67	93.19	96.85	98.32	88.27
$\mathcal{L}_{\mathcal{W}} + \mathcal{L}_{MI} + \mathcal{L}_{\mathcal{GC}}$	3	74.77	96.80	99.11	72.01	83.48	98.96	99.90	86.06	93.15	96.91	98.54	88.32

Table 3: Ablation study on SYSU-MM01 and RegDB datasets. 2 channels represents the model only adopts RGB and IR images as inputs, while 3 channels means that we also add another auxiliary modality (Ye et al. 2021) as input.

memory-based cross-modality mutual information constraint $\mathcal{L}_{MI}$; 3) the global modality-agnostic cluster contrast objective $\mathcal{L}_{GC}$. To real how each ingredient contributes to the performance improvement, we conduct comprehensive ablation study to analyze different elements in Eq. 10.

Specifically, we implement five variants of the proposed method as follows: 1) "CE": Using only the cross-entropy loss to train network as one baseline; 2) $\mathcal{L}^u$: Using only the

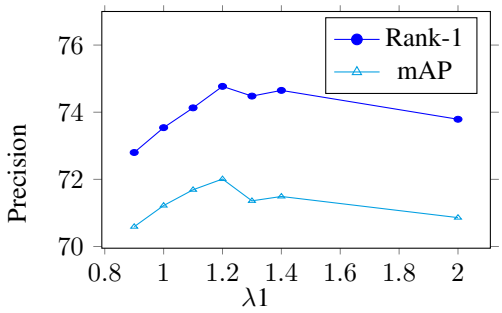

Figure 2: Rank-1 and mAP accuracies with varying values of λ_1 on SYSU-MM01 under Single-shot All-search mode.

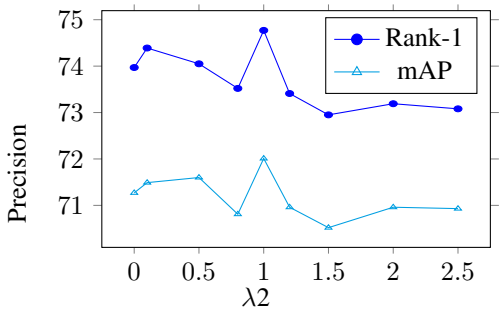

Figure 3: Rank-1 and mAP accuracies with varying values of λ_2 on SYSU-MM01 under Single-shot All-search mode.

modality-agnostic cluster contrast objective as illustrated in the last term of Eq. 4; 3) $\mathcal{L}^{\mathcal{W}}$: Training the network with objective function $\mathcal{L}_{\mathcal{W}}$; 4) $\mathcal{L}_{\mathcal{W}} + \mathcal{L}_{MI}$: Training the network jointly with these two items $\mathcal{L}_{\mathcal{W}}$ and $\mathcal{L}_{MI}$; 5) $\mathcal{L}_{\mathcal{W}} + \mathcal{L}_{MI} + \mathcal{L}_{GC}$: Our final algorithm which optimizes the network jointly with these three items. In order to illustrate the effectiveness of the proposed historical memory-based training mechanism, we also implement their corresponding up-to-date dynamic memory-based mutual information constraint (denoted as $\mathcal{L}^D_{MI}$) and global modality-agnostic cluster contrast (denoted as $\mathcal{L}^D_{GC}$), which means that the cluster centroids used in Eq. 8 and Eq. 9 are from the centroids stored in up-to-date memory bank. Corresponding, we also implement another two variants of the method: 6) $\mathcal{L}_{\mathcal{W}} + \mathcal{L}^D_{MI}$: Training the network joint with $\mathcal{L}_{\mathcal{W}}$ and $\mathcal{L}^D_{MI}$; 7)$\mathcal{L}_{\mathcal{W}} + \mathcal{L}^D_{MI} + \mathcal{L}^D_{GC}$: The dynamically trained network with the overall objective.

The performances of these method variants are summarized in Table 3, where the ablation experiments are conducted on both SYSU-MM01 and RegDB datasets under three experiment settings. By comparing the performance of methods "CE" and $\mathcal{L}^u$, we can conclude that the memory-bank based cluster contrast learning is better than that of the baseline cross-entropy method. When we add the mutual information constraint $\mathcal{L}_{MI}$ and the global cluster contrast objective $\mathcal{L}_{GC}$ into the baseline method $\mathcal{L}_{\mathcal{W}}$ step by step, the performance under all the experiment settings could be further improved gradually. Specifically, comparing the methods $\mathcal{L}_{\mathcal{W}} + \mathcal{L}_{MI}$ with $\mathcal{L}_{\mathcal{W}} + \mathcal{L}^D_{MI}$, as well as $\mathcal{L}_{\mathcal{W}} + \mathcal{L}_{MI} + \mathcal{L}_{GC}$ with $\mathcal{L}_{\mathcal{W}} + \mathcal{L}^D_{MI} + \mathcal{L}^D_{GC}$, we can clearly conclude the superiority of our proposed training mechanism by simultaneously utilize both historical and up-to-date learned proxies for enhanced cross-modality association, which improves the corresponding baseline method by an average margin of 2.68% mAP under All-search Single-shot experiment setting, and 2.48% mAP under Indoor-search Single-shot setting, on SYSU-MM01 dataset.

As we implement our method based on existing channel augmented method (Ye et al. 2021) which manually generates color-irrelevant images as the auxiliary modality, we also implement another baseline version of the proposed method (2 channels) without using auxiliary modality data. Experiment results shown in Table 3 also illustrate the effectiveness of the proposed method.

Hyper-Parameter Sensitivity Analysis As is shown in Figure 2 and Figure 3. As defined in Eq. 10, the overall objective function contains two hyper-parameters (i.e., λ_1 and λ_2) to balance the following three components: $\mathcal{L}_{\mathcal{W}}$, $\mathcal{L}_{MI}$ and $\mathcal{L}_{GC}$. To investigate the effect of hyper-parameters on the model performance, we conduct comprehensive experiments with various values of these parameters. Figure 2 illustrates the model performance in terms of Rank-1 accuracy and mAP with varying values of λ_1 from 0.8 to 2.0 when λ_2 is set to 1.0, on SYSU-MM01 dataset under All-search Single-shot experiment setting. We can clearly see that the model could obtain best performance when $\lambda_1 = 1.2$. Correspondingly, Figure 3 shows model performance with varying values of λ_2 from 0.0 to 2.5 when $\lambda_1 = 1.2$, under the same experiment setting as that in Figure 2. It can be seen that our method yields best performance when $\lambda_1 = 1.2$ and $\lambda_2 = 1.0$.

Conclusion

In this paper, we propose a aggregated memory-based cross-modality deep metric learning for VI-ReID. The proposed method can not only benefit from the increasing number of learned modality-aware and modality-agnostic centroid proxies for cluster contrast and mutual information learning, but also simultaneously utilizes both historical and up-to-date learned cluster proxies to further suppress the modality discrepancy. This training mechanism helps to achieve stronger "pulling close" effect between two modality features. Extensive experiment results demonstrate the superiority of the proposed method. In the future, we would like to further study the memory-based multi-proxy deep metric learning, and extend our work to other cross-modality matching tasks.

Acknowledgements

This work was supported in part by the National Key Research and Development Program of China under Grant 2018AAA0103202, in part by the National Natural Science Foundation of China under Grant 62176198,U22A2096, 62036007, 61922066, 61876142, in part by the Technology Innovation Leading Program of Shaanxi under Grant 2022QFY01-15, in part by Open Research Projects of Zhejiang Lab under Grant 2021KG0AB01, in part by the Natural Science Basic Research Plan in Shaanxi Province of China under Grant 2021JQ-198.

References

Chen, Y.; Wan, L.; Li, Z.; Jing, Q.; and Sun, Z. 2021. Neural feature search for rgb-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 587–597.

Cheng, D.; Gong, Y.; Zhou, S.; Wang, J.; and Zheng, N. 2016. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *Proceedings of the iEEE conference on computer vision and pattern recognition*, 1335–1344.

Cheng, D.; Zhou, J.; Wang, N.; and Gao, X. 2022. Hybrid Dynamic Contrast and Probability Distillation for Unsupervised Person Re-Id. *IEEE Transactions on Image Processing*, 31: 3334–3346.

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.; Liu, X.; Zhu, S.; and Tan, P. 2021. Cluster contrast for unsupervised person re-identification. *arXiv preprint arXiv:2103.11568*.

Deng, J.; Guo, J.; Yang, J.; Lattas, A.; and Zafeiriou, S. 2021. Variational prototype learning for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11906–11915.

Fu, C.; Hu, Y.; Wu, X.; Shi, H.; Mei, T.; and He, R. 2021. CM-NAS: Cross-modality neural architecture search for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11823–11832.

Hao, Y.; Wang, N.; Li, J.; and Gao, X. 2019. HSME: Hypersphere manifold embedding for visible thermal person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 8385–8392.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Li, D.; Wei, X.; Hong, X.; and Gong, Y. 2020. Infrared-visible cross-modal person re-identification with an x modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 4610–4617.

Liu, J.; Sun, Y.; Zhu, F.; Pei, H.; Yang, Y.; and Li, W. 2022. Learning Memory-Augmented Unidirectional Metrics for Cross-Modality Person Re-Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19366–19375.

Lu, Y.; Wu, Y.; Liu, B.; Zhang, T.; Li, B.; Chu, Q.; and Yu, N. 2020. Cross-modality person re-identification with shared-specific feature transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13379–13389.

Nguyen, D. T.; Hong, H. G.; Kim, K. W.; and Park, K. R. 2017. Person recognition system based on a combination of body images from visible light and thermal cameras. volume 17, 605. MDPI.

Wang, G.; Zhang, T.; Cheng, J.; Liu, S.; Yang, Y.; and Hou, Z. 2019a. RGB-infrared cross-modality person re-identification via joint pixel and feature alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3623–3632.

Wang, X.; Girshick, R.; Gupta, A.; and He, K. 2018. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7794–7803.

Wang, X.; Zhang, H.; Huang, W.; and Scott, M. R. 2020. Cross-batch memory for embedding learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6388–6397.

Wang, Z.; Wang, Z.; Zheng, Y.; Chuang, Y.-Y.; and Satoh, S. 2019b. Learning to reduce dual-level discrepancy for infrared-visible person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 618–626.

Wei, Z.; Yang, X.; Wang, N.; and Gao, X. 2021. Syncretic modality collaborative learning for visible infrared person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 225–234.

Wu, A.; Zheng, W.-S.; Yu, H.-X.; Gong, S.; and Lai, J. 2017. RGB-infrared cross-modality person re-identification. In *Proceedings of the IEEE international conference on computer vision*, 5380–5389.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4330–4339.

Yao, H.; and Xu, C. 2021. Dual Cluster Contrastive learning for Object Re-Identification. *arXiv preprint arXiv:2112.04662*.

Ye, M.; Lan, X.; Li, J.; and Yuen, P. 2018. Hierarchical discriminative learning for visible thermal person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Ye, M.; Ruan, W.; Du, B.; and Shou, M. Z. 2021. Channel augmented joint learning for visible-infrared recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13567–13576.

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 *European Conference on Computer Vision*, 229–247. Springer.

Ye, M.; Shen, J.; and Shao, L. 2020. Visible-infrared person re-identification via homogeneous augmented tri-modal learning. volume 16, 728–739. IEEE.

Zhang, Q.; Lai, C.; Liu, J.; Huang, N.; and Han, J. 2022. FMCNet: Feature-Level Modality Compensation for Visible-Infrared Person Re-Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7349–7358.