Syncretic Modality Collaborative Learning for Visible Infrared Person Re-Identification

Ziyu Wei[1], Xi Yang[1*], Nannan Wang[1], Xinbo Gao[2]
[1]The State Key Laboratory of Integrated Services Networks, Xidian University
[2]Key Laboratory of Image Cognition, Chongqing University of Posts and Telecommunications
zywei_xd@stu.xidian.edu.cn, {yangx,nnwang}@xidian.edu.cn, gaoxb@cqupt.edu.cn

Abstract

Visible infrared person re-identification (VI-REID) aims to match pedestrian images between the daytime visible and nighttime infrared camera views. The large cross-modality discrepancies have become the bottleneck which limits the performance of VI-REID. Existing methods mainly focus on capturing cross-modality sharable representations by learning an identity classifier. However, the heterogeneous pedestrian images taken by different spectrum cameras differ significantly in image styles, resulting in inferior discriminability of feature representations. To alleviate the above problem, this paper explores the correlation between two modalities and proposes a novel syncretic modality collaborative learning (SMCL) model to bridge the cross-modality gap. A new modality that incorporates features of heterogeneous images is constructed automatically to steer the generation of modality-invariant representations. Challenge enhanced homogeneity learning (CEHL) and auxiliary distributional similarity learning (ADSL) are integrated to project heterogeneous features on a unified space and enlarge the inter-class disparity, thus strengthening the discriminative power. Extensive experiments on two cross-modality benchmarks demonstrate the effectiveness and superiority of the proposed method. Especially, on SYSU-MM01 dataset, our SMCL model achieves 67.39% rank-1 accuracy and 61.78% mAP, surpassing the cutting-edge works by a large margin.

1. Introduction

Person re-identification (Re-ID) plays an essential role in video surveillance, which automatically searches person images across multiple non-overlapping cameras [39, 36]. Recently, fast-growing works contribute to visible modality person Re-ID and have achieved remarkable performance [1, 28]. However, visible cameras cannot capture enough identity information in the dark. To ensure the safety of pedestrians at night, infrared cameras are deployed to acquire infrared person images, cooperating with visible cameras for 24-hour video surveillance. Hence, visible infrared person re-identification (VI-REID) [27] has emerged to retrieve visible (infrared) images according to the given infrared (visible) images.

VI-REID is challenging due to the considerable visual differences among heterogeneous pedestrian images. Existing studies aim to address this challenge mainly from two aspects, *i.e.*, image-level and feature-level. To achieve modality unification, image generation-based methods [11, 22, 25, 3, 23] are proposed to translate heterogeneous images to the same modality for style consistency. However, the introduction of additional noise during the image translation affects the extraction of discriminative features. To ensure feature alignment, dual-path networks are exploited to obtain modality-specific and modality-invariant representations [38, 33, 7, 5, 32], but the last few layers are difficult to map the specific representations of each modality to a shared space. Subsequently, one-stream weight-sharing network is introduced in massive works [4, 24, 30, 9] to directly extract modality-sharable features. However, the performance of these methods is far inferior to that of visible modality person Re-ID because of the significant color discrepancies between two modalities.

Recently, several researches have built a new modality and combine with two real modalities to conduct tri-modal sharable feature learning, which gains inspiring performance. Li *et al.* [13] introduce an auxiliary X modality as an assistant for modality-invariant feature generation. Ye *et al.* [37] propose a homogeneous augmented grayscale modality and enhance the robustness against color variations. However, these methods neglect the distribution of features in the intermediate modality. As shown in Figure 1, since the images of X modality and grayscale modality are directly generated by the visible images without considering the infrared images, there are two main drawbacks encountering in feature distribution of testing set: 1) The

[*] Corresponding author

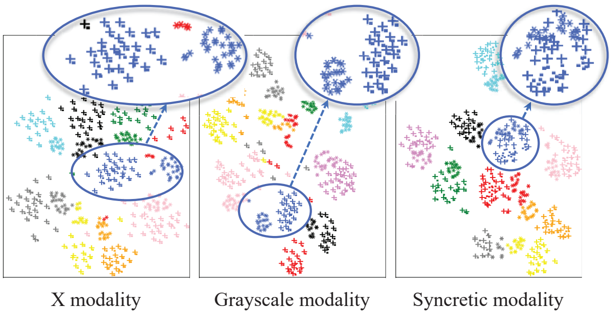


Figure 1. Feature representation distribution of ten randomly selected identities in visible, infrared and three kinds of auxiliary modalities. "+" means visible modality, "*" means infrared modality, and "·" means the auxiliary X, grayscale or the syncretic modality. Different colors represent different identities.

feature distribution of intermediate modality is highly correlated with that of visible modality but not related to infrared modality. 2) The representations of infrared images and visible images still have a great gap in the learned embedding space. Therefore, it is difficult to establish distribution correlation among three modalities only based on the transitional images generated by visible images.

To break above limitations, this paper introduces a novel syncretic modality collaborative learning (SMCL) model to improve the similarity of feature distributions among heterogeneous pedestrian images. The images of syncretic modality are self-generated from both visible and infrared images, thus reserving their common representations. Mutual interaction of three modalities prompts the generation of modality-shared pedestrian features. Specifically, since the infrared images lack color information and are hard to differentiate, we perform challenge enhanced homogeneity learning (CEHL) by bringing a pressure to bear on the identity classifier of infrared images, thus reinforcing the discriminative capability of embedded representations. Furthermore, auxiliary distributional similarity learning (ADSL) is designed to minimize the distance between the centers of data distribution via three directional constraints. Ultimately, we introduce incremental training (IT) strategy by firstly conducting representation learning to roughly restrict the cross-modality feature distribution and then executing metric learning to further narrow-down the modality discrepancies.

Through the proposed method, the feature distributions of visible, infrared and syncretic modalities are visualized in Figure 1. Compared with X modality and grayscale modality, the feature distribution of our syncretic modality is separated from that of visible modality, and the generated images no longer correspond to the visible images one-to-one. Consequently, the syncretic modality really works in feature learning and metric learning. In addition, under the guidance of syncretic modality, the representations of heterogeneous images with the same identity have been assembled, and the feature distances of different identities have been enlarged, thus promoting the performance of VI-REID. The experimental results on SYSU-MM01 and RegDB datasets validate the effectiveness of our method.

The main contributions of this paper can be summarized as follows:

- We propose a novel syncretic modality collaborative learning model for VI-REID task by constructing a self-generated modality which combines visible and infrared image information. Joint learning of three modalities induces the network to capture modality-invariant representations with high discriminability.

- We introduce challenge enhanced homogeneity learning to increase the difficulty of infrared image classifier, thereby urging the network to obtain more discriminative features for correct classification. Besides, auxiliary distributional similarity learning is employed to shrink the cross-modality gap through tri-directional distance suppression.

- We develop incremental training scheme to handle the distribution of heterogeneous images from coarse to fine, thus learning more effective modality-shared discriminative representations for VI-REID. The performance of our SMCL model outperforms the state-of-the-art methods by a remarkable margin.

2. Related Work

Visible modality person Re-ID. The considerable viewpoint changes, human posture variations and resolution changes under different visible cameras are the main challenges of visible modality person Re-ID. The improvements of this task in existing deep learning-based methods are mostly from two aspects, *i.e.*, representation learning [18, 10, 15] and metric learning [14, 20, 29]. In representation learning-based methods, ID-discriminative embedding (IDE) model [40] is usually introduced to regard each person identity as a unique class for image classification. Sun *et al.* [21] partition the person features into p horizontal stripes and input each part of feature into a classifier to extract fine-grained representations. Liu *et al.* [15] propose a view confusion mechanism to learn view-invariant representations. In metric learning-based methods, discriminative features are learned by narrowing down the feature distances between person images. Triplet loss [8] and its improved version [2] are exploited for metric learning. Ye *et al.* [36] exploit weighted regularization triplet loss to optimize the distance between positive and negative samples without fixed margin. Song *et al.* [20] improve the triplet loss and propose the lifted structured embedding by comparing the positive pair with all negative pairs. However, above methods are only exploited for single-modality person Re-ID. There are no specific designs for cross-modality pedestrian retrieval.

Visible infrared person Re-ID. The large disparity of visible and infrared images makes VI-REID a challenging task. Wu *et al.* [27] firstly define the VI-REID problem and contribute a new multiple modality Re-ID dataset SYSU-MM01 for research. Simultaneously, they design a one-stream deep zero-padding framework to explore domain-specific structure automatically in network. Afterwards, Ye *et al.* [38] exploit bi-directional top-ranking loss to handle the modality gap. Dai *et al.* [4] introduce a one-stream network with adversarial learning to compete with representation learning for performance improvement. Moreover, Ye *et al.* [35] propose dynamic dual-attentive aggregation (DDAG) learning to mine both intra-modality part-level and cross-modality graph-level contextual cues for VI-REID. However, the cross-modality representations in above-mentioned methods are hard to map into the consistent space, thus limiting the performance.

To mitigate the modality gap, massive image translation-based methods are developed to firstly achieve modality unification and then learn modality-shared representations. D^{2}RL [25] utilizes variational autoencoders to convert images between visible and infrared modalities. Then, they combine arbitrary modality's person images and the generated heterogeneous images as multi-spectral images to reduce the appearance discrepancy. Wang *et al.* [22] only perform unidirectional translation from visible to infrared modality, and conduct representation learning with the real infrared images and fake infrared images which generated from RGB images. Hi-CMD [3] attempts to capture ID-discriminative and color-irrelevant representations for cross-modality person retrieval. In addition, an auxiliary X modality [13] and the grayscale augmented modality [37] are proposed to better bridge the modality gap with tri-modal learning. However, the learned representations of these two self-generated modalities are close to the visible modality data but far away from the infrared modality data, thus affecting the ability of metric learning.

3. Proposed Method

In this section, we introduce the details of the proposed syncretic modality collaborative learning (SMCL) model for VI-REID. As shown in Figure 2, we firstly propose the **syncretic modality generative module (SMGM)** with a lightweight network, and then exploit **challenge enhanced homogeneity learning (CEHL)** to acquire modality-shared representations. In addition, **auxiliary distributional similarity learning (ADSL)** is employed to narrow-down the cross-modality gap. Finally, an **incremental training (IT) strategy** is introduced to constrain the feature distribution of heterogeneous images from coarse to fine.

3.1. Syncretic Modality Generative Module

In this section, we construct the self-generated syncretic modality which is significant in the subsequent representation learning and metric learning. First, we denote the input visible images and infrared images as $\{v_n|v_n \in V\}_{n=1}^{N}$ and $\{i_n|i_n \in I\}_{n=1}^{N}$, where N is the number of visible and infrared images in a mini-batch. The heterogeneous images are sent into a lightweight network composed of two 1×1 convolutional layers. Specially, we conduct a pixel-to-pixel feature fusion operation to build the syncretic modality after the first convolutional layer, which can be expressed as:

$$\boldsymbol{S}_n = \boldsymbol{V}_n \odot \boldsymbol{I}_n, n \in [1, N], \tag{1}$$

where the feature maps $\boldsymbol{S} \in \mathbb{R}^{C\times H\times W}$, $\boldsymbol{V} \in \mathbb{R}^{C\times H\times W}$, $\boldsymbol{I} \in \mathbb{R}^{C\times H\times W}$, C is the total number of channels and $H \times W$ is the feature map size. "$\odot$" represents the Hadamard product operation. Then, a ReLU activation layer [12] is provided to improve the non-linear ability of the syncretic modality representations. With the second 1×1 convolutional operation, the feature size of syncretic modality is consistent with that of infrared and visible images, so that they can be sent to the parameter-sharing CNN for tri-modality sharable feature learning. The images of the constructed syncretic modality maintain spatial information and pedestrian structure information. Importantly, they reserve the representations of visible and infrared images, rather than only visible images in X-modality [13] and grayscale modality-based methods [37].

3.2. Challenge Enhanced Homogeneity Learning

To acquire modality-sharable identity-discriminative representations, we introduce CEHL to project cross-modality representations on a consistent space. Through CNN, global average pooling (GAP) and batch normalization (BN) operations, the feature vectors are fed into fully connected layers for identity classification. Softmax loss is usually utilized in most person Re-ID methods for discriminative representation learning. Since the visible images and syncretic images have rich color information, the softmax loss for visible representations can be defined as:

$$\mathcal{L}_{id}^{V} = -\frac{1}{N}\sum_{n=1}^{N} \log \frac{e^{\boldsymbol{W}_{y_n}^{T}\boldsymbol{f}_n^{V}}}{\sum_{u=1}^{U} e^{\boldsymbol{W}_{u}^{T}\boldsymbol{f}_n^{V}}}, \tag{2}$$

where y_n and $\boldsymbol{f}_n^V$ are the identity and feature vector of n-th pedestrian image, N is the number of visible images in a mini-batch, U is the number of identities, and $\boldsymbol{W}_u$ is the classifier for u-th identity. The softmax loss for syncretic features $\mathcal{L}_{id}^{S}$ can be denoted by the same form. With the supervision of softmax loss, the network can learn salient modality-invariant representations from visible and syncretic images.

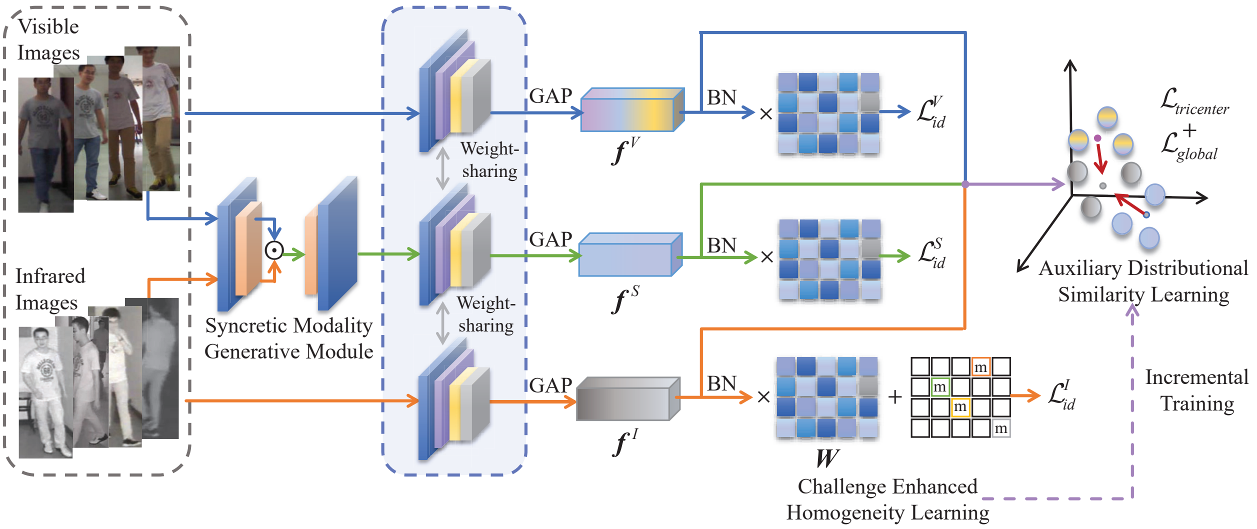


Figure 2. The proposed SMCL model for VI-REID which contains syncretic modality generative module, challenge enhanced homogeneity learning, auxiliary distributional similarity learning and incremental training strategy. The syncretic features generated via the syncretic modality generative module are exploited with visible and infrared images for modality-sharable feature learning. For CEHL, the improved identity losses ($\mathcal{L}_{id}^V, \mathcal{L}_{id}^S, \mathcal{L}_{id}^I$) are leveraged to enhance the discriminative power of the embedding features. For ADSL, tri-directional center-based constrained loss ($\mathcal{L}_{tricenter}$) and global center-constrained loss ($\mathcal{L}_{global}$) are integrated to handle the cross-modality gaps. Finally, IT strategy is conducted to constrain the feature distribution from coarse to fine and improve the training efficiency.

However, the key challenge of VI-REID mainly lies in the lack of homogeneous representations between infrared and visible images. The classifier with standard softmax loss has weak discriminative power for infrared images . To enhance the capability of the identity classifier, we increase the degree of difficulty to the classifier and design an improved softmax loss which can be formulated as:

$$\mathcal{L}_{id}^I = -\frac{1}{N}\sum_{n=1}^{N}\log\frac{e^{\boldsymbol{W}_{y_n}^T \boldsymbol{f}_n^I - m}}{\sum_{u=1}^{U} e^{\boldsymbol{W}_u^T \boldsymbol{f}_n^I}}, \tag{3}$$

where m is the degree of difficulty. The manual pressure stimulates the network to further learn identity-specific features for correct classification. Meanwhile, the joint of syncretic modality in the training phase brings more modality-shared information, thereby boosting the intra-class cross-modality similarity. The overall identity loss in challenge enhanced homogeneity learning can be written as:

$$\mathcal{L}_{id} = \mathcal{L}_{id}^V + \mathcal{L}_{id}^S + \mathcal{L}_{id}^I. \tag{4}$$

3.3. Auxiliary Distributional Similarity Learning

To enhance the cross-modality intra-class similarity and enlarge the intra-modality inter-class disparity, we consider the correlation of three modalities and design a tri-directional center-based constrained loss and a global center-constrained loss. We leverage the center of feature distribution in syncretic modality as an anchor. As shown in Figure 3, suppose that there are $P \times K$ images of P identities in a mini-batch, where each identity contains K images. The feature distributional center of an identity in syncretic modality can be expressed as:

$$\boldsymbol{c}_s^p = \frac{1}{K}\sum_{k=1}^{K}\boldsymbol{s}_k^p, p \in [1, P], \tag{5}$$

where $\boldsymbol{s}_k^p$ is the feature vector of k-th image output from GAP. We introduce a tri-directional center-based constrained loss to handle the distances between the anchor and centers of other modalities, which can be interpreted as:

$$\begin{aligned}\mathcal{L}_{tricenter} = &\sum_{p=1}^{P}\max[(\rho + d(\boldsymbol{c}_s^p, \boldsymbol{c}_v^p) - \min_{p\neq j} d(\boldsymbol{c}_s^p, \boldsymbol{c}_s^j)), 0]\\ &+ \sum_{p=1}^{P}\max[(\rho + d(\boldsymbol{c}_s^p, \boldsymbol{c}_i^p) - \min_{p\neq j} d(\boldsymbol{c}_s^p, \boldsymbol{c}_s^j)), 0],\end{aligned} \tag{6}$$

where $\boldsymbol{c}_v^p$ and $\boldsymbol{c}_i^p$ are the centers of visible and infrared features for the p-th identity, p and j represent different identities within a mini-batch. $d(\cdot)$ denotes the Euclidean distance between two centers. We aim to pull close the distances between centers of different modalities for the same identity and push away the centers of syncretic modality for different identities, thus suppressing cross-modality variations while ensuring high discriminability.

Moreover, to avoid falling into local optimum with the center of syncretic modality as an anchor, we exploit a

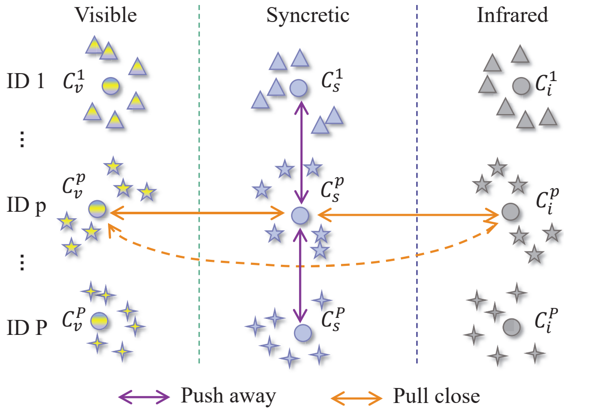


Figure 3. Illustration of the auxiliary distributional similarity learning which contains tri-directional center-based constrained loss (solid line) and a global center-constrained loss (dotted line). Different colors and geometric shapes denote different modalities and identities, respectively. The circle represents the center of feature distribution of a modality for an identity.

global center-constrained loss to directly restrain the distance of centers between visible and infrared features, which can be formulated as:

$$\mathcal{L}_{global} = \sum_{p=1}^{P} \|\boldsymbol{c}_v^p - \boldsymbol{c}_i^p\|_2 . \tag{7}$$

For the features of the same identity, we not only regard the features of syncretic modality as an intermediary to promote the cross-modality distributional similarity, but also increase a straightforward restriction for heterogeneous images; for the features of different identities, the centers of the syncretic modality are utilized to enlarge the feature distance. The overall loss in ADSL can be written as:

$$\mathcal{L}_{adsl} = \mathcal{L}_{tricenter} + \mathcal{L}_{global}. \tag{8}$$

The total loss of our SMCL model can be denoted as:

$$\mathcal{L}_{total} = \mathcal{L}_{id} + \lambda \mathcal{L}_{adsl}. \tag{9}$$

3.4. Incremental Training Strategy

Most of person Re-ID methods jointly exploit representation learning and metric learning to obtain effective features for person matching. However, the heterogeneous images have random distribution in the initial state. The joint training may cause inconsistency in the direction of gradient descent for the two learning manners, thus affecting the training efficiency. To improve the training efficiency and optimize the objective function to the maximum extent, we propose an incremental training (IT) scheme as shown in Algorithm 1. The CEHL performed in the initial stage of training coarsely clusters the features of the same pedestrian, and the subsequent collaborative learning of CEHL and ADSL narrows the feature distance and reinforces the similarity of cross-modality intra-class representations. The proposed IT strategy can handle the distribution of heterogeneous images from coarse to fine, thus enhancing the discriminability of the embedding features.

Algorithm 1 Incremental Training of SMCL Model

```
Input: Visible image set $V = \{v_1, ..., v_n\}$, infrared image set $I = \{i_1, ..., i_n\}$, label set $Y = \{y_1, ..., y_n\}$, the total training epoch $T$, the start epoch of collaborative learning $Q$, parameters $m$, $\rho$ and $\lambda$;
for $t = 1$ to $T$ do
    Generate syncretic feature maps $\boldsymbol{S}$ by Eq.(1)
    Output $\boldsymbol{f}^V$, $\boldsymbol{f}^S$ and $\boldsymbol{f}^I$ from the backbone
    Compute the identity loss $\mathcal{L}_{id}$ by Eq.(4)
    if $t < Q$ then
        Update parameters $\theta_{id}$ of CEHL
    else
        Calculate $\mathcal{L}_{adsl}$ according to Eq.(8)
        Calculate $\mathcal{L}_{total}$ according to Eq.(9)
        Update parameters $\theta_{id}$ of CEHL
        Update parameters $\theta_{adsl}$ of ADSL
    end if
end for
Output: Optimized model of the proposed method
```

4. Experiments

4.1. Datasets and Settings

Datasets. To evaluate the performance of the proposed method, we conduct experiments on two public cross-modality person Re-ID datasets, *i.e.*, SYSU-MM01 [27] and RegDB [19]. SYSU-MM01 [27] consists of 44,745 heterogeneous pedestrian images of 491 identities captured by 4 visible cameras and 2 infrared cameras. There are 22,258 visible images and 11,909 infrared images of 395 identities in the training set. In the testing phase, infrared and visible images are adopted as query set and gallery set, respectively. The search mode consists of all-search mode and indoor-search mode. For both modes, we adopt single-shot and multi-shot settings to evaluate the performance. RegDB [19] contains 4120 images of 412 identities acquired by dual-camera systems. Each person includes 10 visible images and 10 thermal images. We follow the evaluation protocol in [38]. To achieve statistically stable results, the procedure is repeated for 10 trials to calculate the average performance. The standard Cumulative Matching Characteristics (CMC) curve and mean Average Precision (mAP) are adopted as the evaluation metrics.

Implementation details. The proposed method is implemented with PyTorch framework on two TITAN RTX GPUs. We adopt ResNet-50 model pretrained on ImageNet

Table 1. Different components of the proposed method on two datasets. CMC (%) at rank 1 and mAP (%).

B	CEHL	ADSL	IT	SYSU-MM01		RegDB	
				r=1	mAP	r=1	mAP
✓	×	×	×	57.00	55.49	75.77	70.52
✓	✓	×	×	59.97	56.01	77.52	73.40
✓	✓	✓	×	63.16	59.92	80.63	73.85
✓	✓	✓	✓	67.39	61.78	83.05	78.57

Table 2. Performance of different auxiliary modalities on two datasets. CMC (%) at rank 1 and mAP (%).

Auxiliary Modality	SYSU-MM01		RegDB	
	r=1	mAP	r=1	mAP
Nothing	57.58	54.69	78.49	76.89
X modality	62.34	59.33	79.46	73.00
grayscale modality	64.23	60.88	74.36	69.41
syncretic modality	67.39	61.78	83.05	78.57

as the backbone network and modify the stride of the last convolutional block to 1. For a training batch, we randomly select heterogeneous images of 4 person. Each identity contains 4 infrared images and 4 visible images. All the images are resized to $3 \times 384 \times 128$. Adam optimizer is exploited with the base learning rate initialized to 3.5×10^{-4}, and then decayed to 3.5×10^{-5}, 3.5×10^{-6} after 40, 70 epochs, respectively. We exploit CEHL and introduce ADSL for collaborative learning after 220 epochs with totally 300 epochs on SYSU-MM01 dataset. For RegDB dataset, we train the model with totally 200 epochs and introduce ADSL after 120 epochs. The parameter λ and ρ are set to 0.5 and 0.3, respectively. During the testing phase, we utilize cosine similarity to measure the distances of heterogeneous features.

4.2. Ablation Study

Effectiveness of each component. We evaluate the performance of each component on SYSU-MM01 and RegDB datasets in Table 1. Compared with the baseline model (B) which utilizes the SMGM and standard softmax loss for representation learning, the mAP of CEHL is enhanced by 0.52% and 2.88% on SYSU-MM01 and RegDB datasets, respectively. Hence, the degree of difficulty m can effectively promote the discriminative feature learning. When we perform CEHL and ADSL from scratch simultaneously, the rank-1 accuracy and mAP are improved by 3.19% and 3.91% on SYSU-MM01 dataset. Therefore, ADSL can further shrink the cross-modality discrepancies and strengthen the discriminative power of the network. Ultimately, after introducing the IT strategy, we achieve the highest mAP of 61.78% and 78.57% on two datasets, which indicates the effectiveness of our IT strategy.

Effectiveness of the syncretic modality. To verify the superiority of the proposed syncretic modality, we replace syncretic modality with X modality [13] and grayscale modality [37] which are generated from visible images. In addition, "Nothing" means that the input of CNN is visible and infrared images, without images from other auxiliary modalities. The comparison results are reported in Table 2. The mAP of our method without auxiliary modality is at least 4.64% lower than that with auxiliary modality on SYSU-MM01 dataset. Consequently, the auxiliary modality can induce the generation of modality-shared representations. On SYSU-MM01 dataset, the performance of grayscale modality is higher than that of X modality, which proves that the images of grayscale modality assist the network to map more heterogeneous features on the consistent space compared with the images of X modality. On the contrary, X modality is more effective than grayscale modality for RegDB dataset. The proposed method with syncretic modality improves rank-1 accuracy by 3.16% as compared to that with grayscale modality on SUSU-MM01 dataset, and boosts the mAP by 5.57% compared with X modality on RegDB dataset. Therefore, our syncretic modality can effectively combine visible and infrared images for modality-sharable representation learning.

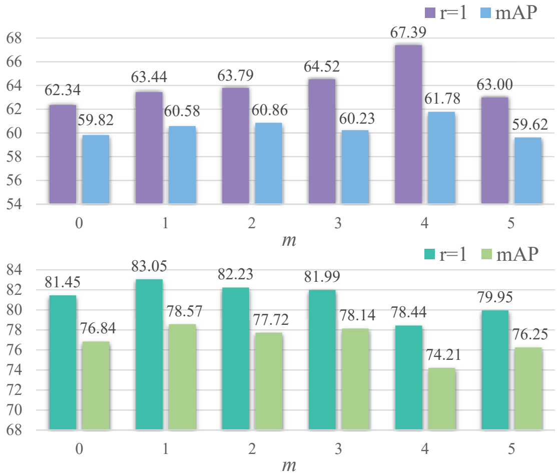


Figure 4. Comparison of different m in CEHL on SYSU-MM01 (top row) and RegDB dataset (bottom row). CMC (%) at rank 1 and mAP (%).

Evaluation of different margin m. The margin in the proposed CEHL affects the difficulty of classification in representation learning. We vary m from 0 to 5 and report the performance comparison on two datasets in Figure 4 . For SYSU-MM01 dataset, we achieve the highest mAP and rank-1 accuracy when m is set to 4. Since the pedestrian images on SYSU-MM01 dataset have great intra-modality and cross-modality divergences caused by illumination and body posture, it is necessary to increase the classification difficulty of identity classifier to facilitate the discriminative feature learning. For RegDB dataset, the heterogeneous pedestrian images taken by binocular cameras have minor intra-class difference. Therefore, favorable perfor-

Table 3. Comparison of computational costs in metric learning.

	positive	negative
L_{bh_tri}	$2PK \times (2K-1)$	$2PK \times 2(P-1)K$
L_{bdtr}	$2PK \times K + 2PK \times (K-1)$	$2PK \times 2(P-1)K$
$L_{bicenter}$	$2PK$	$2PK \times (P-1)$
L_{hc_tri}	$2P$	$2P \times 2(P-1)$
L_{adsl}	$2P + P$	$P \times (P-1)$

Table 4. Comparison of rank-1 (%), mAP (%) and training time (s) in metric learning.

	rank-1	mAP	Training Time
L_{bh_tri}	61.53	59.13	0.35
L_{bdtr}	59.08	56.63	0.35
$L_{bicenter}$	60.50	57.09	0.34
L_{hc_tri}	61.29	57.86	**0.33**
L_{adsl}	**67.39**	**61.78**	**0.33**

mance can be obtained with a slight increase of the classifier difficulty, and the performance will drop with the large degree of difficulty. We finally set m to 1 on RegDB dataset.

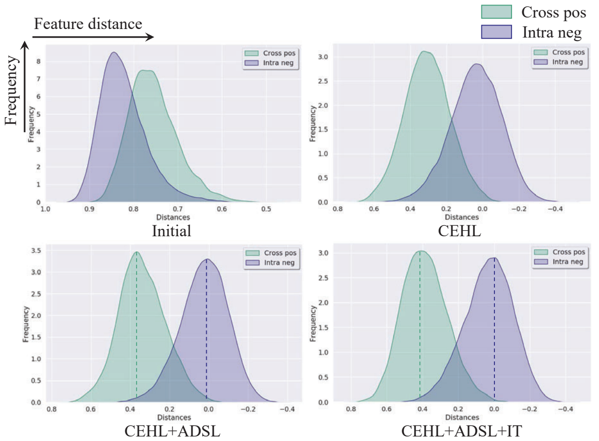
4.3. Computational Cost in Metric Learning

We compare the computational cost of our ADSL with other metric learning methods for VI-REID. L_{bh_tri} represents the batch hard sampling utilized in most VI-REID methods [4, 31, 36] . L_{bdtr} means dual-constrained top-ranking loss which is leveraged in BDTR [38] and HSME [7]. We also compare $L_{bicenter}$ in the sample to center-based method [34] and L_{hc_tri} in center to center-based method [16]. Suppose that a mini-batch consists of $P \times K$ images of P identities, the computational costs are shown in Table 3. For sample to sample-based methods, L_{bdtr} constrains the cross-modality and intra-modality discrepancies simultaneously, which has inspiring performance but high computational consumption. $L_{bicenter}$ and L_{hc_tri} don't require to compute distance between sample and sample, thus reducing the computational cost greatly. In our ADSL method, the feature centers of syncretic modality are viewed as the anchor. Hence, the improved triplet loss L_{adsl} is designed to calculate the pairwise distance between feature centers of syncretic modality for different identities, which achieves the lowest computational cost compared to above methods. Moreover, as shown in Table 4, due to the fast matrix operation of GPU, the training time of ADSL and other methods are similar. However, the performance of ours outperforms theirs to a great extent, which validates the effectiveness of our method.

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

In this section, we compare our method with cutting-edge VI-REID methods on two public datasets.

Comparison on SYSU-MM01 dataset. Our model achieves 67.39% rank-1 and 61.78% mAP on SYSU-MM01 dataset. As shown in Table 5, for feature learning-based methods with one-stream network ([27], [4], [30], [26], [6], [36], [13], [32], [37]), the proposed method exhibits inspiring performance, which outperforms them at least 12.1% in rank-1 accuracy and 7.89% in mAP under all-search single-shot mode. Therefore, our SMCL model can capture more modality-shared and discriminative features than other one-stream network-based methods. Furthermore, compared with two-stream network-based methods ([33], [38], [34], [7], [31], [35]), our method exceeds DDAG by 12.64% in rank-1 and 8.76% in mAP. Specially, for the best former method cm-SSFT, we compare its performance with single query (SQ) which is widely used in most methods. The rank-1 accuracy and mAP of our method are 19.69% and 7.68% higher than cm-SSFT, respectively. Besides, SMCL also improves the rank-1 by 5.79% compared to it in all queries (AQ) search mode, which verifies the superiority of the proposed method. For those image generation-based methods ([25], [3], [23], [22]), our syncretic modality generated from lightweight network can effectively maps heterogeneous images on a common space, so the performance of ours surpasses theirs by a large margin.



Initial CEHL CEHL+ADSL CEHL+ADSL+IT

Figure 5. The distribution of the cosine distance between cross-modality positive samples and intra-modality negative samples.

Comparison on RegDB dataset. To prove the effectiveness and robustness of our method, we conduct experiments on different query settings to compare with the state-of-the-art methods in Table 6. Under visible to thermal query settings, our method is 9.46% and 4.54% higher than the best former method SIM [9] in rank-1 accuracy and mAP. Moreover, the improvement in rank-1 and map is 7.81% and 0.27% on thermal to visible query setting, respectively. Hence, our SMCL model is robust against different query settings and can better narrow the feature distribution of heterogeneous images.

4.5. Visualization Analysis

We visualize the cosine distance distribution of cross-modality positive samples and intra-modality negative sam-

Table 5. Comparison with state-of-the-art methods on SYSU-MM01 datasets. CMC (%) at rank r and mAP (%).

Methods		All-search								Indoor-search							
		Single-shot				Multi-shot				Single-shot				Multi-shot			
		r=1	r=10	r=20	mAP	r=1	r=10	r=20	mAP	r=1	r=10	r=20	mAP	r=1	r=10	r=20	mAP
Zero-Padding [27]		14.80	54.12	71.33	15.95	19.13	61.40	78.41	10.89	20.58	68.38	85.79	26.92	24.43	75.86	91.32	18.64
TONE [33]		12.52	50.72	68.60	14.42	-	-	-	-	-	-	-	-	-	-	-	-
BDTR [38]		17.01	55.43	71.96	19.66	-	-	-	-	-	-	-	-	-	-	-	-
eBDTR [34]		27.82	67.34	81.34	28.42	-	-	-	-	32.46	77.42	89.62	42.46	-	-	-	-
D-HSME [7]		20.68	62.74	77.95	23.12	-	-	-	-	-	-	-	-	-	-	-	-
cmGAN [4]		26.97	67.51	80.56	27.80	31.49	72.74	85.01	22.27	31.63	77.23	89.18	42.19	37.00	80.94	92.11	32.76
D^2RL [25]		28.90	70.60	82.40	29.20	-	-	-	-	-	-	-	-	-	-	-	-
MAC [31]		33.26	79.04	90.09	36.22	-	-	-	-	33.37	82.49	93.69	44.95	-	-	-	-
Hi-CMD [3]		34.94	77.58	-	35.94	-	-	-	-	-	-	-	-	-	-	-	-
JSIA-ReID [23]		38.1	80.7	89.9	36.9	45.1	85.7	93.8	29.5	43.8	86.2	94.2	52.9	52.7	91.1	96.4	42.7
expAT Loss [30]		38.57	76.64	86.39	38.61	44.71	69.82	77.87	32.20	-	-	-	-	-	-	-	-
AlignGAN [22]		42.4	85.0	93.7	40.7	51.5	89.4	95.7	33.9	45.9	87.6	94.4	54.3	57.1	92.7	97.4	45.3
FMSP [26]		43.56	74.61	86.25	44.98	-	-	-	-	48.62	79.01	89.50	57.50	-	-	-	-
DFE [6]		48.71	88.86	95.27	48.59	54.63	91.62	96.83	42.14	52.25	89.86	95.85	59.68	59.62	94.45	98.07	50.60
AGW [36]		47.50	-	-	47.65	-	-	-	-	54.17	-	-	62.97	-	-	-	-
XIV-ReID [13]		49.92	89.79	95.96	50.73	-	-	-	-	-	-	-	-	-	-	-	-
MACE [32]		51.64	87.25	94.44	50.11	-	-	-	-	57.35	93.02	97.47	64.79	-	-	-	-
DDAG [35]		54.75	90.39	95.81	53.02	-	-	-	-	61.02	94.06	98.41	67.98	-	-	-	-
HAT [37]		55.29	92.14	97.36	53.89	-	-	-	-	62.10	95.75	99.20	69.37	-	-	-	-
cm-SSFT [17]	SQ	47.7	-	-	54.1	57.4	-	-	59.1	-	-	-	-	-	-	-	-
	AQ	61.6	89.2	93.9	63.2	63.4	91.2	95.7	62.0	70.5	94.9	97.7	72.6	73.0	96.3	99.1	72.4
Ours (SMCL)		**67.39**	**92.87**	96.76	**61.78**	**72.15**	90.66	94.32	54.93	**68.84**	**96.55**	98.77	**75.56**	**79.57**	**95.33**	98.00	**66.57**

Table 6. Comparison with state-of-the-art methods on RegDB datasets. CMC (%) at rank 1 and mAP (%).

Methods		Visible to Thermal		Thermal to Visible	
		r=1	mAP	r=1	mAP
Zero-Padding [27]		17.75	18.90	16.63	17.82
TONE [33]		16.87	14.92	13.86	16.98
BDTR [38]		33.47	31.83	32.72	31.10
eBDTR [34]		34.62	33.46	34.21	32.49
D-HSME [7]		50.85	47.00	50.15	46.46
D^2RL [25]		43.4	44.1	-	-
MAC [31]		36.43	37.03	36.20	36.63
Hi-CMD [3]		70.93	66.04	-	-
JSIA-ReID [23]		48.5	49.3	48.1	48.9
expAT Loss [30]		66.48	67.31	67.45	66.51
AlignGAN [22]		57.9	53.6	56.3	53.4
FMSP [26]		65.07	64.50	-	-
DFE [6]		70.13	69.14	67.99	66.70
AGW [36]		70.05	66.37	-	-
XIV-ReID [13]		62.21	60.18	-	-
MACE [32]		72.37	69.09	72.12	68.57
DDAG [35]		69.34	63.46	68.06	61.80
SIM [9]		74.47	75.29	75.24	78.30
HAT [37]		71.83	67.56	70.02	66.30
cm-SSFT [17]	SQ	65.4	65.6	63.8	64.2
	AQ	72.3	72.9	71.0	71.7
Ours (SMCL)		**83.93**	**79.83**	**83.05**	**78.57**

ples with different components of our SMCL model. As shown in Figure 5, the differences between negative samples are less than that of positive samples in the initial state. With the addition of CEHL and ADSL, the disparity of intra-modality negative samples gradually becomes greater than that of cross-modality positive samples. After introducing the IT strategy, there is a slight increase in the distance between two distributions. Therefore, our method can effectively enlarge the distance between negative samples and reduce the discrepancies between positive samples, thereby improving the retrieval accuracy.

5. Conclusion

In this paper, we propose a novel syncretic modality collaborative learning (SMCL) model to learn modality-invariant identity-discriminative representations for VI-REID. The self-generated features of syncretic modality reserve the significant information of visible and infrared images, which can steer the network to project heterogeneous images on a common space with the challenge enhanced homogeneity learning and auxiliary distributional similarity learning. Massive experiments on SYSU-MM01 and RegDB datasets demonstrate the superior performance of our SMCL model.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant nos. 61976166, 62036007, 61772402, 62050175, 61922066, and 61876142, in part by the Key Research and Development Program of Shaanxi under Grant 2021GY-030, in part by the Innovation Capacity Support Plan of Shaanxi Province under Grant 2020KJXX-027, in part by the Young Talent Fund of University Association for Science and Technology, Shaanxi, China, under Grant 20180104, in part by the Fundamental Research Funds for the Central Universities under Grant JB210115, and in part by the Innovation Fund of Xidian University.

References

[1] Tianlong Chen, Shaojin Ding, Jingyi Xie, Ye Yuan, Wuyang Chen, Yang Yang, Zhou Ren, and Zhangyang Wang. ABD-Net: Attentive but diverse person re-identification. In *ICCV*, 2019.

[2] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *CVPR*, pages 403–412, 2017.

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

[4] Pingyang Dai, Rongrong Ji, Haibin Wang, Qiong Wu, and Yuyu Huang. Cross-modality person re-identification with generative adversarial training. In *IJCAI*, volume 1, page 2, 2018.

[5] Zhanxiang Feng, Jianhuang Lai, and Xiaohua Xie. Learning modality-specific representations for visible-infrared person re-identification. *IEEE Transactions on Image Processing*, 29:579–590, 2019.

[6] Yi Hao, Nannan Wang, Xinbo Gao, Jie Li, and Xiaoyu Wang. Dual-alignment feature embedding for cross-modality person re-identification. In *ACM MM*, pages 57–65, 2019.

[7] Yi Hao, Nannan Wang, Jie Li, and Xinbo Gao. Hsme: Hypersphere manifold embedding for visible thermal person re-identification. In *AAAI*, volume 33, pages 8385–8392, 2019.

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

[9] Mengxi Jia, Yunpeng Zhai, Shijian Lu, Siwei Ma, and Jian Zhang. A similarity inference metric for rgb-infrared cross-modality person re-identification. In *IJCAI*, pages 1026–1032, 2020.

[10] Jiening Jiao, Wei-Shi Zheng, Ancong Wu, Xiatian Zhu, and Shaogang Gong. Deep low-resolution person re-identification. In *AAAI*, volume 32, 2018.

[11] Vladimir V Kniaz, Vladimir A Knyaz, Jiri Hladuvka, Walter G Kropatsch, and Vladimir Mizginov. Thermalgan: Multimodal color-to-thermal image translation for person re-identification in multispectral dataset. In *ECCV*, 2018.

[12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[13] Diangang Li, Xing Wei, Xiaopeng Hong, and Yihong Gong. Infrared-visible cross-modal person re-identification with an x modality. In *AAAI*, volume 34, pages 4610–4617, 2020.

[14] Shengcai Liao and Stan Z Li. Efficient psd constrained asymmetric metric learning for person re-identification. In *ICCV*, pages 3685–3693, 2015.

[15] Fangyi Liu and Lei Zhang. View confusion feature learning for person re-identification. In *ICCV*, pages 6639–6648, 2019.

[16] Haijun Liu, Xiaoheng Tan, and Xichuan Zhou. Parameter sharing exploration and hetero-center triplet loss for visible-thermal person re-identification. *IEEE Transactions on Multimedia*, pages 1–1, 2020.

[17] Yan Lu, Yue Wu, Bin Liu, Tianzhu Zhang, Baopu Li, Qi Chu, and Nenghai Yu. Cross-modality person re-identification with shared-specific feature transfer. In *CVPR*, 2020.

[18] Tetsu Matsukawa, Takahiro Okabe, Einoshin Suzuki, and Yoichi Sato. Hierarchical gaussian descriptor for person re-identification. In *CVPR*, pages 1363–1372, 2016.

[19] 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.

[20] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *CVPR*, pages 4004–4012, 2016.

[21] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, pages 480–496, 2018.

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

[23] Guan-An Wang, Tianzhu Zhang, Yang Yang, Jian Cheng, Jianlong Chang, Xu Liang, and Zeng-Guang Hou. Cross-modality paired-images generation for rgb-infrared person re-identification. In *AAAI*, volume 34, pages 12144–12151, 2020.

[24] Pingyu Wang, Zhicheng Zhao, Fei Su, Yanyun Zhao, Haiying Wang, Lei Yang, and Yang Li. Deep multi-patch matching network for visible thermal person re-identification. *IEEE Transactions on Multimedia*, 2020.

[25] Zhixiang Wang, Zheng Wang, Yinqiang Zheng, Yung-Yu Chuang, and Shin'ichi Satoh. Learning to reduce dual-level discrepancy for infrared-visible person re-identification. In *CVPR*, pages 618–626, 2019.

[26] Ancong Wu, Wei-Shi Zheng, Shaogang Gong, and Jianhuang Lai. Rgb-ir person re-identification by cross-modality similarity preservation. *International journal of computer vision*, 128(6):1765–1785, 2020.

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

[28] Xi Yang, Liangchen Liu, Nannan Wang, and Xinbo Gao. A two-stream dynamic pyramid representation model for video-based person re-identification. *IEEE Transactions on Image Processing*, 30:6266–6276, 2021.

[29] Xun Yang, Peicheng Zhou, and Meng Wang. Person reidentification via structural deep metric learning. *IEEE Transactions on Neural Networks and Learning Systems*, 30(10):2987–2998, 2018.

[30] Hanrong Ye, Hong Liu, Fanyang Meng, and Xia Li. Bidirectional exponential angular triplet loss for rgb-infrared person re-identification. *IEEE Transactions on Image Processing*, 30:1583–1595, 2020.

[31] Mang Ye, Xiangyuan Lan, and Qingming Leng. Modality-aware collaborative learning for visible thermal person re-identification. In *ACM MM*, pages 347–355, 2019.

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

[33] Mang Ye, Xiangyuan Lan, Jiawei Li, and Pong Yuen. Hierarchical discriminative learning for visible thermal person re-identification. In *AAAI*, volume 32, 2018.

[34] Mang Ye, Xiangyuan Lan, Zheng Wang, and Pong C Yuen. Bi-directional center-constrained top-ranking for visible thermal person re-identification. *IEEE Transactions on Information Forensics and Security*, 15:407–419, 2019.

[35] 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.

[36] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

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

[38] Mang Ye, Zheng Wang, Xiangyuan Lan, and Pong C Yuen. Visible thermal person re-identification via dual-constrained top-ranking. In *IJCAI*, volume 1, page 2, 2018.

[39] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016.

[40] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. Person re-identification in the wild. In *CVPR*, pages 1367–1376, 2017.