

X modality

Grayscale modality

Syncretic modality



[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

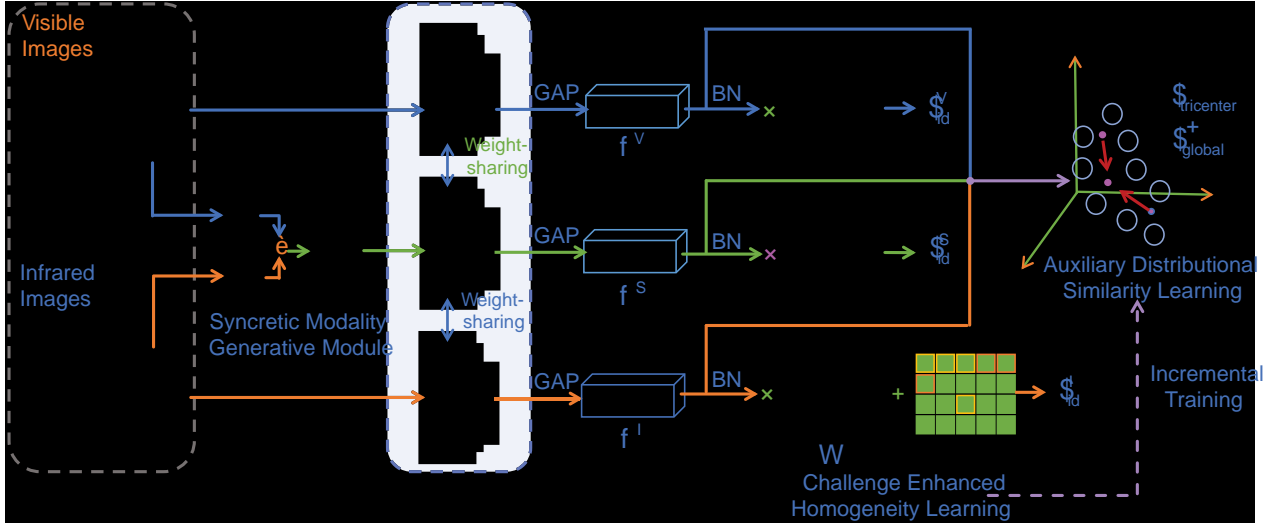


Figure 2. The proposed SMCL model for VI-REID which contains syncretic modality generative module, challenge enhanced homogeneity learning, auxiliary distributional similarity learning and incremental training strategy. The syncretic features generated via the syncretic modality generative module are exploited with visible and infrared images for modality-sharable feature learning. For CEHL, the improved identity losses $(L_{id}^v, L_{id}^s, L_{id}^l)$ are leveraged to enhance the discriminative power of the embedding features. For ADSL, tri-directional center-based constrained loss $(L_{tricenter})$ and global center-constrained loss (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:

$$L_{id}^l = \sum_{n=1}^N \frac{1}{N} \log \frac{e^{W_{yn}^T f_n^l \tilde{S}_m}}{\sum_{u=1}^U e^{W_{un}^T f_n^l}}, \quad (3)$$

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:

$$L_{id} = L_{id}^v + L_{id}^s + L_{id}^l. \quad (4)$$

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 K images of P identities in a mini-batch, where each identity contains m images.

$$c_s^p = \frac{1}{K} \sum_{k=1}^K s_k^p, p \in [1, P], \quad (5)$$

where 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:

$$L_{tricenter} = \sum_{p=1}^P \max[(\alpha + d(c_s^p, c_v^p) \tilde{S} \min_{p=j} d(c_s^p, c_s^j)), 0] + \sum_{p=1}^P \max[(\alpha + d(c_s^p, c_i^p) \tilde{S} \min_{p=j} d(c_s^p, c_s^j)), 0], \quad (6)$$

where c_v^p and 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

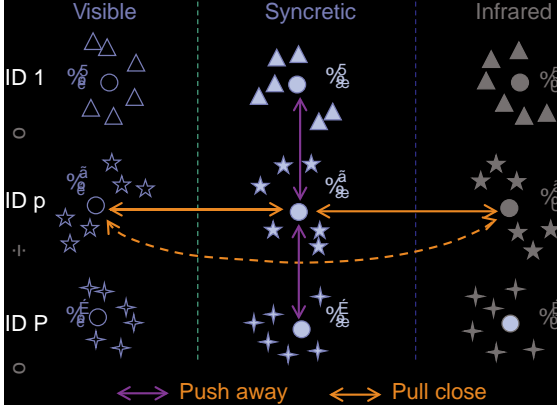


Figure 3. Illustration of the auxiliary distributional similarity learning which contains tri-directional center-based constrained loss (solid line) and a global center-constrained loss (dotted line). Different colors and geometric shapes denote different modalities and identities, respectively. The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

$$L_{\text{global}} = \sum_{p=1}^P c_v^p \tilde{S} c_i^p \quad (7)$$

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:

$$L_{\text{adsl}} = L_{\text{tricenter}} + L_{\text{global}} \quad (8)$$

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

$$L_{\text{total}} = L_{\text{id}} + L_{\text{adsl}} \quad (9)$$

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 results, the procedure is repeated for 10 trials to calculate 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

Algorithm 1 Incremental Training of SMCL Model

Input: Visible image set $V = \{v_1, \dots, v_n\}$, infrared image set $I = \{i_1, \dots, i_n\}$, label set $Y = \{y_1, \dots, y_n\}$, the total training epoch T , the start epoch of collaborative learning Q , parameters θ , and ϕ ;

- 1: for $t = 1$ to T do
- 2: Generate syncretic feature maps by Eq.(1)
- 3: Output f^v, f^s and f^i from the backbone
- 4: Compute the identity loss L_{id} by Eq.(4)
- 5: if $t < Q$ then
- 6: Update parameters θ of CEHL
- 7: else
- 8: Calculate L_{adsl} according to Eq.(8)
- 9: Calculate L_{total} according to Eq.(9)
- 10: Update parameters θ of CEHL
- 11: Update parameters ϕ of ADSL
- 12: end if
- 13: end for

Output: Optimized model of the proposed method

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.

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, 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

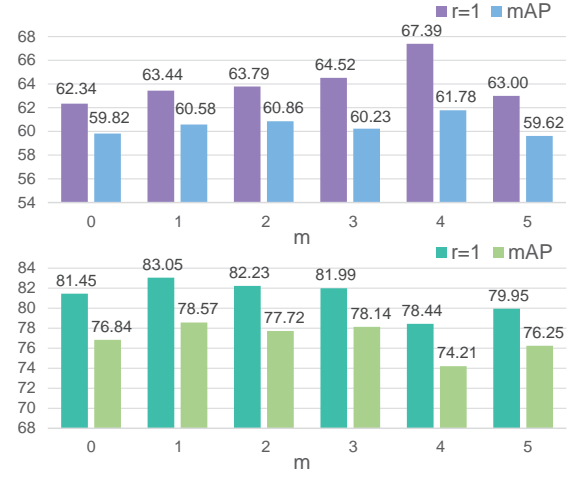


Figure 4. Comparison of different margin m in CEHL on SYSU-MM01 (top row) and RegDB dataset (bottom row). CMC (%) at rank 1 and mAP (%).

addition, •NothingŽ 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.

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-

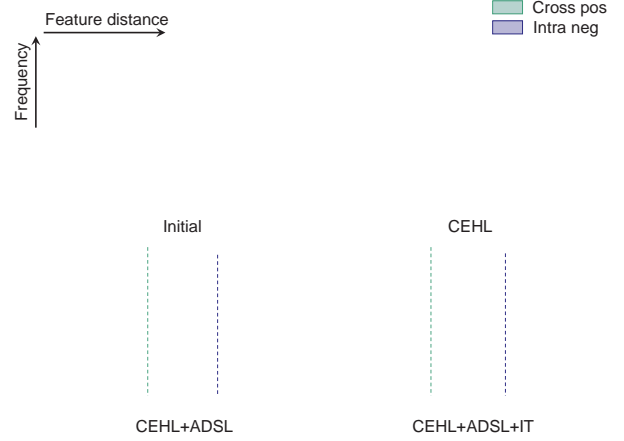


Figure 5. The distribution of the cosine distance between cross-modality positive samples and intra-modality negative samples.

[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.

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-

