Bidirectional Multi-Step Domain Generalization for Visible-Infrared Person Re-Identification: Supplementary

[1]Mahdi Alehdaghi, [2]Pourya Shamsolmoali, [1]Rafael M. O. Cruz, and [1]Eric Granger
[1]LIVIA, ILLS, Dept. of Systems Engineering, ETS Montreal, Canada
[2]Dept. of Computer Science, University of York, UK
mahdi.alehdaghi.1@ens.etsmtl.ca, pshams55@gmail.com,
{rafael.menelau-cruz, eric.granger}@etsmtl.ca

Appendix A. Proofs

In Section 3 of the manuscript, our model seeks to represent input images using discriminative prototypes that are complementary. To ensure their complementarity, we minimize the mutual information (MI) between the data distributions of each pair of prototypes using a contrastive loss between feature representations of the different prototypes. To improve the discrimination, the prototype representations encode ID-related information by maximizing the MI between the joint distribution among all prototypes in the label distribution space. In this section, we prove that by minimizing the cross-entropy loss between features of each prototype class and the person ID in images, we can learn discriminative prototype representations.

A.1. Maximizing $\text{MI}(P^1, \dots, P^K; Y)$

This section provides a proof that $\text{MI}(P^1, \dots, P^K; Y)$ (Eq. 10) can be lower-bounded by :

$$\text{MI}(P^1; Y) + \cdots + \text{MI}(P^K; Y), \tag{A.1}$$

following the properties of mutual information.

P.1 (Nonnegativity) For every pair of random variables X and Y:

$$\text{MI}(X; Y) \geq 0 \tag{A.2}$$

P.2 For every pair random variables X, Y that are independent:

$$\text{MI}(X; Y) = 0. \tag{A.3}$$

P.3 (Monotonicity) For every three random variables X, Y and Z:

$$\text{MI}(X; Y; Z) \leq \text{MI}(X; Y) \tag{A.4}$$

P.4 For every three random variables X, Y and Z, the mutual information of joint distortions X and Z to Y is:

$$\text{MI}(X, Z; Y) = \text{MI}(X; Y) + \text{MI}(Z; Y) - \text{MI}(X; Z; Y) \tag{A.5}$$

$$MI(P^1, \dots, P^K; Y). \tag{A.6}$$

Theorem 1 *Let $P^1, \dots, P^K$ and Y be random variables with domains $\mathcal{P}^1, \dots, \mathcal{P}^K$ and $\mathcal{Y}$, respectively. Let every pair P^k and P^q ($k \neq q$) be independent. Then, maximizing MI$(P^1, \dots, P^K; Y)$ can be approximated by maximizing the sum of MI between each of P^k to Y, $\sum_{k=1}^{K} MI(P^k; Y)$.*

Proof 1 *First, we define $\tilde{P}^k$ as the joint distribution of $P^{k+1}, \ldots, P^K$. Using the* **P.4** *we have:*

$$MI(P^1, \tilde{P}^1; Y) = \underbrace{MI(P^1; Y)}_{\alpha} + \underbrace{MI(\tilde{P}^1; Y)}_{\beta} - \underbrace{MI(P^1; \tilde{P}^1; Y)}_{\gamma}, \tag{A.7}$$

To maximize this, α and β should be maximized and γ minimized. Given **P.3**, $\min MI(P^1; \tilde{P}^1; Y)$ *can be upper-bounded by* $\min MI(P^1; \tilde{P}^1)$*:*

$$MI(P^1; \tilde{P}^1; Y) \leq MI(P^1; \tilde{P}^1), \tag{A.8}$$

So by minimizing the right term of Eq. A.8, γ is also minimized. We show that $MI(P^1; \tilde{P}^1) = 0$ by expanding $\tilde{P}^1$ to $(P^2, \tilde{P}^2)$:

$$MI(P^1; P^2, \tilde{P}^2) = MI(P^1; P^2) + MI(P^1; \tilde{P}^2) - MI(P^1; P^2; \tilde{P}^2). \tag{A.9}$$

Given **P.1** *and Eq. 3, $MI(P^1; P^2) = 0$ and $MI(P^1; P^2; \tilde{P}^2) \leq MI(P^1; P^2) = 0$ so that:*

$$MI(P^1; P^2, \tilde{P}^2) = MI(P^1; \tilde{P}^2). \tag{A.10}$$

After recursively expanding $\tilde{P}^2$:

$$MI(P^1; P^2, \tilde{P}^2) = 0. \tag{A.11}$$

To maximize β, we can rewrite and expand recursively in Eq. A.7 :

$$MI(\tilde{P}^1; Y) = MI(P^2, \tilde{P}^2; Y) = \underbrace{MI(P^2; Y)}_{\textit{maximizing}} + \underbrace{MI(\tilde{P}^2; Y)}_{\textit{expanding}} - \underbrace{MI(P^2; \tilde{P}^2; Y)}_{0}. \tag{A.12}$$

Therefore, it can be shown that:

$$MI(\tilde{P}^k; Y) = MI(P^{k+1}, \tilde{P}^{k+1}; Y) = MI(P^k; Y) + \cdots + MI(P^K; Y) \quad \forall k \in \{0, \ldots, K-1\}. \tag{A.13}$$

and for $k = 0$, we have:

$$MI(P^1, \ldots, P^K; Y) = \sum_{k=1}^{K} MI(P^k; Y), \tag{A.14}$$

Finally, for maximizing $MI(P^1, \ldots, P^K; Y)$, we need to maximize each $MI(P^k; Y)$ so that each prototype feature contains Id-related information and is complemented. In other words, each prototype seeks to describe the input images from different aspects.

A.2. Maximizing $\mathrm{MI}(P^k; Y)$

In Section 3, the MI between the representation of each prototype and the label of persons are maximized by minimizing cross-entropy loss (see Eq: 11 of the manuscript). This approximation is formulated as **Proposition 1**.

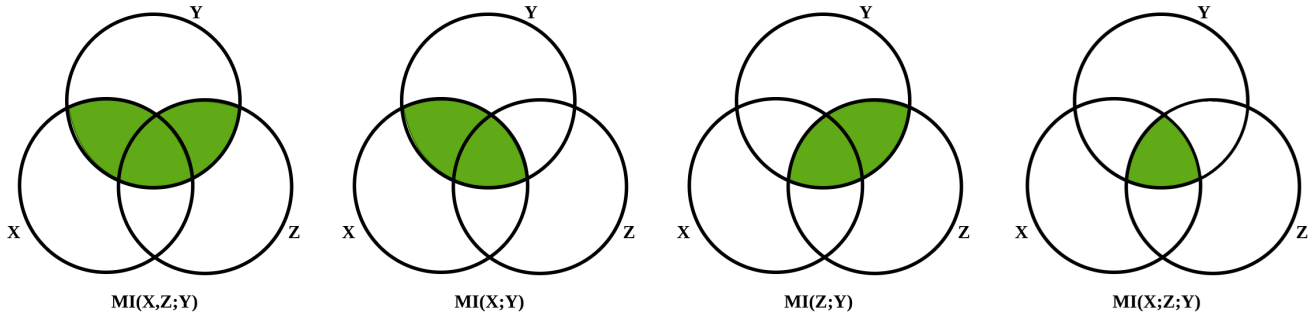


Figure A.1. Venn diagram of theoretic measures for three variables X, Y, and Z, represented by the lower left, upper, and lower right circles, respectively.

Proposition 1 *Let P^k and Y be random variables with domains $\mathcal{P}^k$ and $\mathcal{Y}$, respectively. Minimizing the conditional cross-entropy loss of predicted label $\hat{Y}$, denoted by $\mathcal{H}(Y;\hat{Y}|P^k)$, is equivalent to maximizing the MI$(P^k;Y)$*

Proof 2 *Let us define the MI as entropy,*

$$MI(P^k;Y) = \underbrace{\mathcal{H}(Y)}_{\delta} - \underbrace{\mathcal{H}(Y|P^k)}_{\xi} \tag{A.15}$$

Since the domain $\mathcal{Y}$ does not change, the entropy of the identity δ term is a constant and can therefore be ignored. Maximizing MI(P^k,Y) can only be achieved by minimizing the ξ term. We show that $\mathcal{H}(Y|P^k)$ is upper-bounded by our cross-entropy loss (Eq. 11), and minimizing such loss results in minimizing the ξ term. By expanding its relation to the cross-entropy [1]:

$$\mathcal{H}(Y;\hat{Y}|P^k) = \mathcal{H}(Y|P^k) + \underbrace{\mathcal{D}_{KL}(Y||\hat{Y}|P^k)}_{\geq 0}, \tag{A.16}$$

where:

$$\mathcal{H}(Y|P^k) \leq \mathcal{H}(Y;\hat{Y}|P^k). \tag{A.17}$$

Through the minimization of Eq. 11, training can naturally be decoupled in 2 steps. First, weights of the prototype module are fixed, and only the classifier parameters (i.e., weight W^k of the fully connected layer) are minimized w.r.t. Eq. A.16. Through this step, $\mathcal{D}_{KL}(Y||\hat{Y}|P^k)$ is minimized by adjusting $\hat{Y}$ while the $\mathcal{H}(Y|P^k)$ does not change. In the second step, the prototype module's weights are minimized w.r.t. $\mathcal{H}(Y|P^k)$, while the classifier parameters W^k are fixed.

Appendix B. Additional Details on the proposed method

B.1. Training Algorithm

To train the model, BMDG uses a batch of data containing N_b person with N_p positive images from the V and I modalities. Algorithm 1 shows the details of the BMGD training strategy for optimizing the feature backbone by gradually increasing mixing prototypes.

At first, the prototype mining module extracts K prototypes from infrared and visible images in lines 4 and 5 . Then, at lines 6 and 7, $\mathcal{G}$ function mixes these prototypes from each modality to create two intermediate features. It is noted that the ratio of mixing gradually increases w.r.t the step number t to create more complex samples. To refine the final feature descriptor for input images, the attentive embedding module, $\mathcal{F}$, is applied to prototypes to leverage the attention between them. At the end of each iteration, the model's parameters will be optimized by minimizing the cross-modality ReID objectives between each modality features vector and its gradually created intermediate.

Algorithm 1 BMDG Training Strategy.

Require: $\mathcal{S} = \{\mathcal{V}, \mathcal{I}\}$ as training data and T, K as hyper-parameters

1: **for** $t = 1, \ldots, T$ **do** ▷ over T steps
2: **while** all batches are not selected **do**
3: $x_v^j, x_i^j \leftarrow$ **batchSampler**(N_b, N_p)
4: extract prototypes $\mathbf{A}_v^j$ and global features $\mathbf{g}_v^j$ from visible images v^j ▷ left-side of Fig. 2(a)
5: extract prototypes $\mathbf{A}_i^j$ and global features $\mathbf{g}_i^j$ from infrared images i^j ▷ left-side of Fig. 2(a)
6: $\mathbf{A}_v^{(t)} \leftarrow \mathcal{G}(\mathbf{A}_v^j, \mathbf{A}_i^j, t)$ ▷ V intermediate by gradually increasing the mixing ratio w.r.t t from I prototypes
7: $\mathbf{A}_i^{(t)} \leftarrow \mathcal{G}(\mathbf{A}_i^j, \mathbf{A}_v^j, t)$ ▷ I intermediate by gradually increasing the mixing ratio w.r.t t from V prototypes
8: **if** $t \leq T$ **then**:
9: $\mathbf{f}_v^{(t)} \leftarrow [\mathcal{F}(\mathbf{A}_v^{(t)}); \mathbf{g}_v]$ ▷ embedding intermediate visible features
10: $\mathbf{f}_i^{(t)} \leftarrow [\mathcal{F}(\mathbf{A}_i^{(t)}); \mathbf{g}_i]$ ▷ embedding intermediate infrared features
11: **else**:
12: $\mathbf{f}_v^{(t)} \leftarrow [\mathcal{F}(\mathbf{A}_v^{j(t)}); \mathbf{g}_i]$
13: $\mathbf{f}_i^{(t)} \leftarrow [\mathcal{F}(\mathbf{A}_i^{(t)}); \mathbf{g}_v]$
14: **end if**
15: update model's parameters by optimizing Eq. 19
16: **end while**
17: **end for**

B.2. Attentive Prototype Embedding

Details of the Attentive Prototype Embedding module are depicted in B.1.

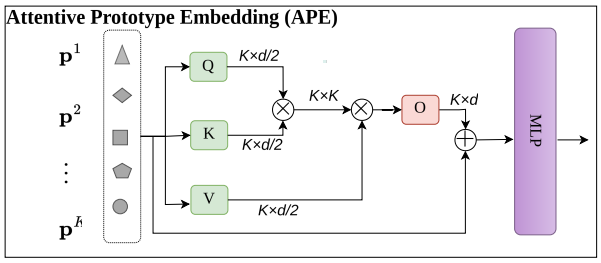


Figure B.1. Attentive prototype embedding (APE) architecture.

Appendix C. Additional Details on the Experimental Methodology

C.1. Datasets:

Research on cross-modal V-I ReID has extensively used the SYSU-MM01 [7], RegDB [6], and recently published LLCM [**?**] datasets. SYSU-MM01 is a large dataset containing more than 22K RGB and 11K IR images of 491 individuals captured with 4 RGB and 2 near-IR cameras, respectively. Of the 491 identities, 395 were dedicated to training, and 96 were dedicated to testing. Depending on the number of images in the gallery, the dataset has two evaluation modes: single-shot and multi-shot. RegDB contains 4,120 co-located V-I images of 412 individuals. Ten trial configurations randomly divide the dataset into two sets of 206 identities for training and testing. The tests are conducted in two ways – comparing I to V (query) and vice versa. An LLCM dataset consists of a large, low-light, cross-modality dataset that is divided into training and testing sets at a 2:1 ratio.

C.2. Experimental protocol:

We used a pre-trained ResNet50 [3] as the deep backbone model. Each batch contains 8 RGB and 8 IR images from 10 randomly selected identities. Each image input is resized to 288 by 144, then cropped and erased randomly, and filled with zero padding or mean pixels. ADAM optimizer with a linear warm-up strategy was used for the optimization process. We trained the model by 180 epochs, in which the initial learning rate is set to 0.0004 and is decreased by factors of 0.1 and 0.01 at 80 and 120 epochs, respectively. $K = 6$, $T = 4$, $\lambda_f = 0.1$, $\lambda_v = 0.05$, $\lambda_p = 0.2$ and $\lambda_i = 0.4$ are set based on the analyses shown in the ablation study in the main paper and in Section D.1. $\lambda_{eq} = 0.5$ is for all experiments.

C.3. Performance measures:

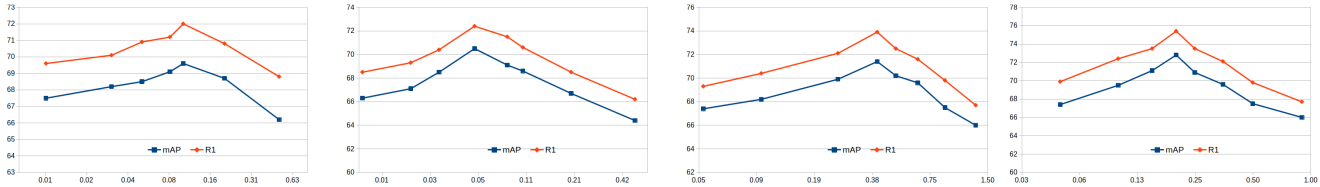
We use Cumulative Matching Characteristics (CMC) and Mean Average Precision (mAP) as assessment metrics in our study. In CMC, rank-k accuracy is measured to determine how likely it is that a precise cross-modality image of the person will be present in the top-k retrieved results. As an alternative, mAP can be used as a measure of image retrieval performance when multiple matching images are found in a gallery.

Appendix D. Additional Quantitative Results

D.1. Hyperparameter values:

This subsection analyzes the impact of λ_f, λ_v, λ_p, and λ_i on V-I ReID accuracy. We initially set $\lambda_v = 0.01$, $\lambda_p = 0.05$, and $\lambda_i = 0.8$, experimenting with various values for λ_f. As shown in Fig. D.1, accuracy improves with increasing λ_f until it reaches 0.1. Elevated λ_f enhances prototype diversity, boosting the discriminative ability of final features in the diverse space. However, excessively high values disperse prototype features in the feature space, diminishing discriminability and hindering accurate identification.

Similar trends are observed when $\lambda_p = 0.05$ and $\lambda_i = 0.8$, varying λ_v from 0.01 to 0.05, resulting in improved performance. Higher λ_v compresses prototype regions excessively, lacking sufficient ID-related information. Conversely, λ_i enhances the discriminative capabilities of prototypes in images. Balancing these factors, we find optimal values of 0.05 and 0.4 for λ_v and λ_i, respectively. Additionally, based on experimentation, we set $\lambda_p = 0.2$ at the end of our analysis.



(a) λ_f. (b) λ_v. (c) λ_i. (d) λ_p.

Figure D.1. Accuracy of the proposed BMDG over λ_f, λ_v, λ_i, and λ_p values on SYSU-MM01 dataset in all-search and single-shot mode.

D.2. Step size and number of part prototypes:

In section 4.2, we discussed the step size and number of part prototypes based on Rank-1 accuracy. Here we report the mAP measurement for Table 4 in paper in Table D.1a and Table **??** in paper in Table D.1b, respectively:

Table D.1. mAP accuracy of BMDG using (a) our prototype mixing and (b) Mixup [9] setting for different numbers of part prototypes (K) and intermediate steps (T).

(a) Prototype exchanging

T	Number of part prototypes (K)					
	3	4	5	6	7	10
0	65.98	67.03	67.23	68.11	67.55	65.66
1	67.42	68.72	69.28	69.46	68.32	69.28
2	69.51	70.08	70.72	71.14	69.97	71.25
3	71.44	71.69	71.82	72.02	71.06	71.67
4	-	71.98	72.15	**72.86**	71.19	69.00
6	-	-	-	72.40	71.17	69.54
10	-	-	-	-	-	69.46

(b) Mixup [9]

T	Number of part prototypes (K)					
	3	4	5	6	7	10
0	65.98	67.03	67.23	68.11	67.55	65.66
1	66.31	67.22	67.31	68.28	68.5	65.8
4	66.50	67.68	67.79	68.52	68.49	65.88
6	66.98	67.77	68.05	**68.73**	68.56	66.02
10	67.04	67.60	67.93	68.65	68.54	65.57

D.3. Model efficiency:

Our BMDG proposed a method to extract alignable part-prototypes in feature extraction and then compute an attention embedding for the final representation features. In Table D.2, we showed each component size and time complexity in the inference time compared to the baseline we used.

Table D.2. Number parameters and floating-point operations at inference time for BMDG and all its sub-modules.

Model	**# of Para. (M)**	**Flops (G)**
Feature Backbone	24.8	5.2
Prototype Mining	3.1	0.2
Attentive Prototype Embedding	1.8	0.3
baseline [8]	24.9	5.2
BMDG	29.7	5.7

Appendix E. Visual Results

E.1. UMAP projections:

To show the effectiveness of BMDG, we randomly select 7 identities from the SYSU-MM01 dataset and project their feature representations using the UMAP method [5] for (a) Baseline, (b) one-step (prototypes without gradual training), and (c) BMDG. Visualization results (Figure E.1) show that compared with the baseline and one-step approach, the feature representations learned with our BMDG method are well clustered according to their respective identity, showing a strong capacity to discriminate. BMDG is effective for learning robust and identity-aware features. Our BMDG method reduces this distance across modalities for each person and provides more separation among samples from different people.

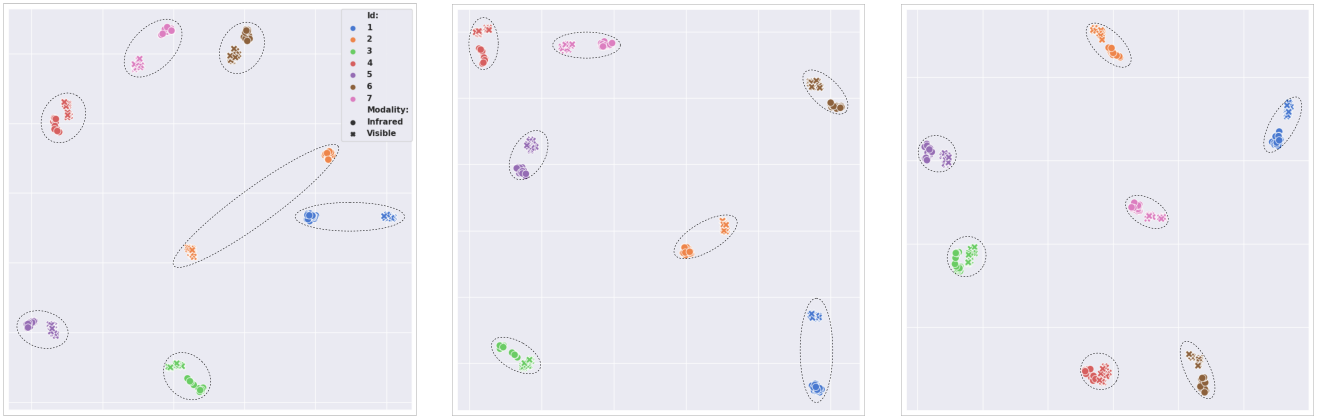
Also, to show how the intermediate features gradually mix the modalities, we draw intermediate features for 6 steps in Figure E.2. At the beginning of training, the features are based on modality while at step 6, the features are concentrated on each identity.

E.2. Domain shift:

To estimate the level of domain shift over data from V and I modalities, we measured the MMD distance for each training epoch. To this end, for each epoch, we selected 10 random images from 50 random identities and extracted the prototype and global features, then measured the MMD distance between the centers of those features for each modality as shown in Fig. E.3(b). We report the normalized MMD distances between I and V features for our BMDG approach when compared with the baseline. Our method reduces this distance more than the y baseline. Thus, the results show that the intermediate domains improve the model robustness to a large multi-modal domain gap by gradually increasing the mix in prototypes over multiple steps.

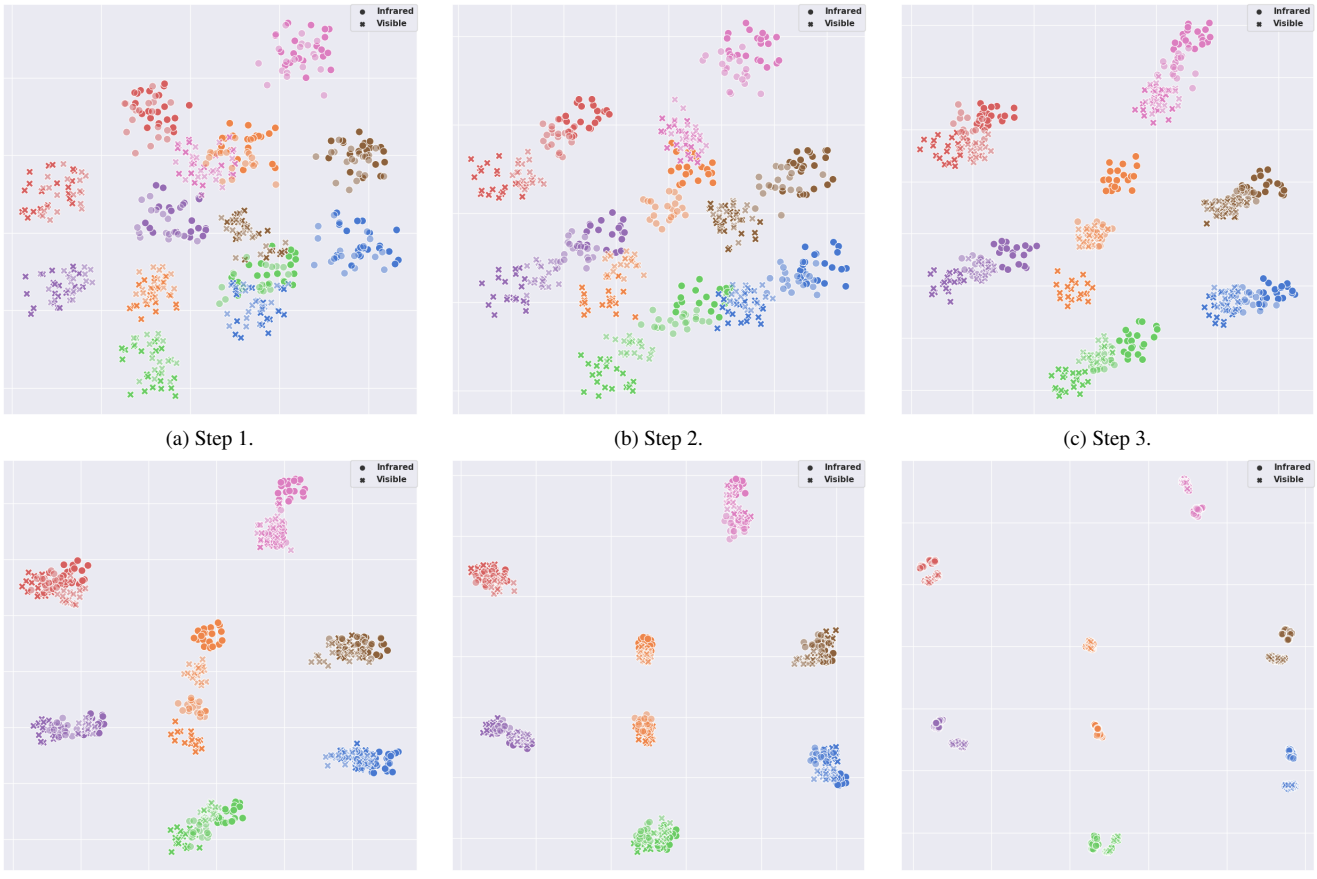
E.3. Part-prototype masking:

To show spatial information related to prototype features, we visualize the score map in the PM module (see Fig. E.4). Our approach encodes prototype regions linked to similar body parts without considering person identity. Our model tries to



(a) Baseline. (b) One Step. (c) BMDG.

Figure E.1. Distributions of learned V and infrared features of 7 identities from SYSU-MM01 dataset for (a) the baseline, (b) one-step using part-prototypes, and (c) our BMDG method by UMAP [5]. Each color shows the identity.

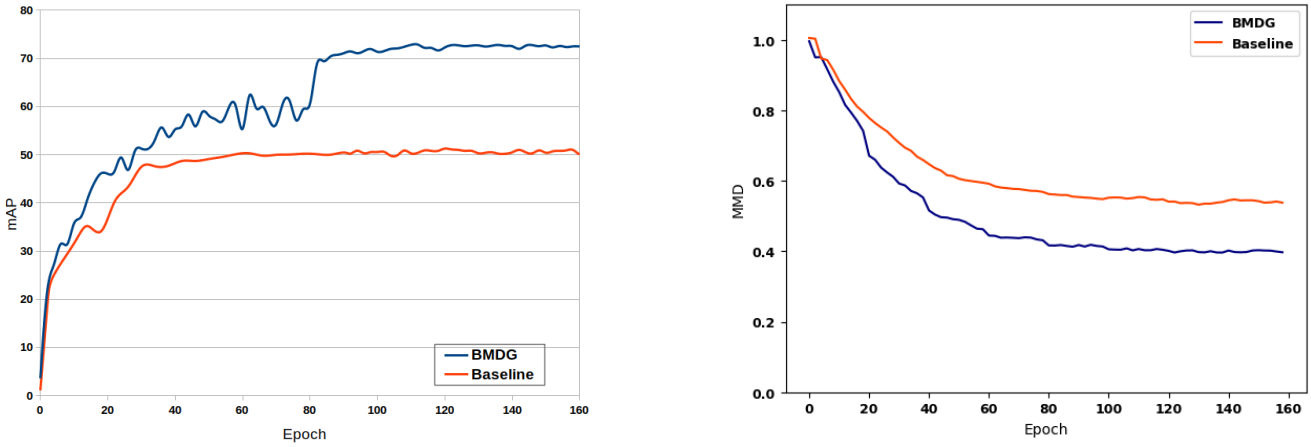


(a) Step 1. (b) Step 2. (c) Step 3.

(d) Step 4. (e) Step 5. (f) Step 6.

Figure E.2. Distributions of learned V and infrared features of 7 identities from SYSU-MM01 dataset in training for 6 steps at epochs of 10,30,50,70,90 and 160 respectively by UMAP [5]. Each color shows the identity. The intermediate features are drawn with lower opacity.

find similar regions for each class of prototypes and then extracts ID-related information for that region. Therefore, BMDG is more robust for matching the same part features.



(a) Learning Curve. (b) BMDG.

Figure E.3. The mAP and domain shift (MMD distance) between I and V modalities over training epochs. (a) The learning curve of BMDG vs Baseline [8]. (b) MMD distance over the center of multiple person's infrared features to visible modality.

(a) Infrared. (b) Visible.

Figure E.4. Prototypes regions extracted by the PRM module for (a) infrared and (b) visible images. Note that the mask size is 18×9, which is then resized to fit the original input image. As shown, the mask of prototypes focuses on similar body parts without accounting for identity.

E.4. Semi-supervised body part detection:

An additional benefit of our HCL module lies in its ability to detect meaningful parts in a semi-supervised manner. By forcing the model to identify semantic regions that are both informative about foreground objects and contrastive to each other, our hierarchical contrastive learning provides robust part detection, even in the absence of part labels. To assess HCL, we fine-tuned our ReID model as a student using a pre-trained part detector [4] on the PASCAL-Part Dataset [2] as the teacher. In Fig. E.5, the results show our model's strong capacity for detecting body parts compared to its teacher.

(a) Infrared. (b) Visible.

Figure E.5. Semi-supervised part discovery on the SYSU-MM01 dataset. The first columns are the (a) infrared and (b) visible images. The second column images are the result from [4], and the last column are results with our fine-tuned model in BMDG.

References

[1] Boudiaf, M., Rony, J., Ziko, I.M., Granger, E., Pedersoli, M., Piantanida, P., Ayed, I.B.: A unifying mutual information view of metric learning: cross-entropy vs. pairwise losses. In: European conference on computer vision. pp. 548–564. Springer (2020) 3

[2] Chen, X., Mottaghi, R., Liu, X., Fidler, S., Urtasun, R., Yuille, A.: Detect what you can: Detecting and representing objects using holistic models and body parts. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1971–1978 (2014) 8

[3] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 5

[4] Li, P., Xu, Y., Wei, Y., Yang, Y.: Self-correction for human parsing. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020). https://doi.org/10.1109/TPAMI.2020.3048039 8, 9

[5] McInnes, L., Healy, J., Melville, J.: Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426 (2018) 6, 7

[6] Nguyen, D.T., Hong, H.G., Kim, K.W., Park, K.R.: Person recognition system based on a combination of body images from visible light and thermal cameras. Sensors **17**(3), 605 (2017) 2, 7, 5

[7] Wu, A., Zheng, W.S., Gong, S., Lai, J.: Rgb-ir person re-identification by cross-modality similarity preservation. International journal of computer vision **128**(6), 1765–1785 (2020) 2, 7, 5

[8] Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., Hoi, S.C.H.: Deep learning for person re-identification: A survey and outlook. arXiv preprint arXiv:2001.04193 (2020) 1, 3, 7, 6, 8

[9] Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. In: International Conference on Learning Representations (2018) 3, 8, 6