



Deposited via The University of York.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/id/eprint/239044/>

Version: Accepted Version

---

**Proceedings Paper:**

Alehdaghi, Mahdi, Shamsolmoali, Pourya, M. O. Cruz, Rafael et al. (2025) Bidirectional Multi-Step Domain Generalization for Visible-Infrared Person Re-Identification. In: IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). , pp. 763-773.

<https://doi.org/10.1109/WACV61041.2025.00084>

---

**Reuse**

This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here:

<https://creativecommons.org/licenses/>

**Takedown**

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing [eprints@whiterose.ac.uk](mailto:eprints@whiterose.ac.uk) including the URL of the record and the reason for the withdrawal request.

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

<sup>1</sup>Mahdi Alehdaghi, <sup>2</sup>Pourya Shamsolmoali, <sup>1</sup>Rafael M. O. Cruz, and <sup>1</sup>Eric Granger

<sup>1</sup>LIVIA, ILLS, Dept. of Systems Engineering, ETS Montreal, Canada

<sup>2</sup>Dept. of Computer Science, University of York, UK

mahdi.alehdaghi.1@ens.etsmtl.ca, pshams55@gmail.com,

{rafael.menelau-cruz, eric.granger}@etsmtl.ca

## Abstract

A key challenge in visible-infrared person re-identification (V-I ReID) is training a backbone model capable of effectively addressing the significant discrepancies across modalities. State-of-the-art methods that generate a single intermediate bridging domain are often less effective, as this generated domain may not adequately capture sufficient common discriminant information. This paper introduces Bidirectional Multi-step Domain Generalization (BMDG), a novel approach for unifying feature representations across diverse modalities. BMDG creates multiple virtual intermediate domains by learning and aligning body part features extracted from both I and V modalities. In particular, our method aims to minimize the cross-modal gap in two steps. First, BMDG aligns modalities in the feature space by learning shared and modality-invariant body part prototypes from V and I images. Then, it generalizes the feature representation by applying bidirectional multi-step learning, which progressively refines feature representations in each step and incorporates more prototypes from both modalities. Based on these prototypes, multiple bridging steps enhance the feature representation. Experiments<sup>1</sup> conducted on V-I ReID datasets indicate that our BMDG approach can outperform state-of-the-art part-based and intermediate generation methods, and can be integrated into other part-based methods to enhance their V-I ReID performance.

## 1. Introduction

V-I ReID is a variant of person ReID that involves matching individuals across RGB and IR cameras. V-I ReID is challenging since it requires matching individuals with significant differences in appearance between V and I modality images. In this context, V-I ReID systems must train a discriminant feature extraction backbone to encode con-

sistent and identifiable attributes in RGB and IR images. Most state-of-the-art methods seek to learn global representations from the whole image by alignment at the image-level [19, 32] or feature-level [28, 44, 47, 59] (see Fig. 1(a)). Other methods extract global modality-invariant features by disentangling them from modality-specific information [7, 42]. However, global feature-based approaches cannot compare local discriminative attributes, resulting in the loss of important discriminant cues about the person. To address this issue and focus on the unique information in different body regions, part-based approaches (see Fig. 1(b)) extract local fine-grained feature representations through horizontal stripes, clustering, or attention mechanisms of spatially extracted features [17, 25, 40, 46]. However, given the domain discrepancies between the V and I modalities and the lack of matching cross-modal parts during training, these methods often learn modality-specific attributes for each part. This typically results in feature representations that are less modality-invariant, and thus provide limited effectiveness for cross-modal matching tasks.

Some methods leverage an intermediate modality to guide the training process such that the cross-modal domain gap is reduced [2, 20, 37, 48, 54] (see Fig. 1(c)). Intermediate modalities may be static images (e.g., gray-scale) [2, 48] or created images from a generative module learned by I and V modalities [20, 37, 54]. The resulting intermediate representations can leverage shared relations between I and V images and improve robustness to modality changes. However, they generate images that tend to contain shared non-discriminative information, resulting in the loss of ID-aware information in the intermediate domain. Moreover, one-step training strategies only create one intermediate space between V and I. This cannot provide sufficient bridging information when modalities have significant discrepancies. Single-step approaches cannot effectively deal with these discrepancies during training. Therefore, we advocate for linking a series of intermediate domains with progressively

<sup>1</sup>Our code is available at: [alehdaghi.github.io/BMDG](https://github.com/alehdaghi/BMDG)

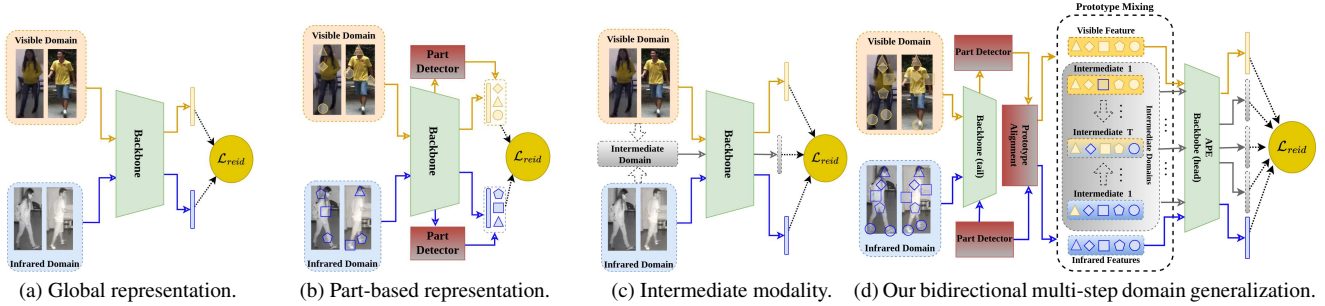


Figure 1. A comparison of training architectures for V-I ReID. Approaches based on (a) a global features representation, and (b) a local part-based representation to preserve locality and global features. (c) Generation of an intermediate bridging domain to guide training. (d) Our BMGD extracts and combines prototypes from modalities in each step to gradually create multiple intermediate bridging domains.

smaller gaps.

To effectively bridge the large cross-modal domain gap while preserving local ID-discriminative information, we introduce Bidirectional Multi-step Domain Generalization (BMDG), a gradual, multi-step, part-based training strategy for person V-I ReID (see Fig. 1(d)). At first, this strategy identifies similar ID-aware part prototypes<sup>2</sup> for each modality, and then it progressively creates intermediate steps by controlling the number of the part prototypes to be mixed. Generated images can become degraded, as discriminative details are lost over consecutive generation steps, making these images less useful for training. Therefore, our proposed BMDG creates multiple intermediate domains in the feature space by aligning, and then combining ID-informative subsets of features from both modalities.

Our BMGD training strategy is comprised of two parts. The prototype alignment module learns to identify and align part-prototype representations that contain discriminative information linked to body parts in I and V images. Then, the bidirectional multi-step learning progressively extracts auxiliary intermediate feature representations by mixing prototypes from both modalities at each step. For ID-aware and robust alignment between modalities, the extracted part-prototypes must meet three conditions: (1) they should be *complementary* to other parts, (2) they should be *interchangeable* and represent consistent parts, and (3) they should be *discriminant* for person matching. Hierarchical contrastive learning is proposed to train the prototype alignment module such that these properties are respected without affecting ReID accuracy. Our novel idea is to progressively create intermediate bridging domains by mixing a growing number of body part prototypes extracted from I and V modalities. This allows leveraging common semantic information between modalities, along with discriminative information among individuals.

Bidirectional multi-step learning relies on auxiliary intermediate feature spaces that are created at each step,

<sup>2</sup>We define a *part prototype* as a feature representation corresponding to a specific body part in a cropped image of a person.

by progressively mixing semantically consistent prototypes from each modality. While prototype-based approaches [12, 24, 46, 58] concentrate solely on enhancing discriminatory representation using local descriptors, BMDG aligns and mixes local part information in a gradual training process to mitigate domain discrepancies within the representation space. Unlike the Part-Mix method [18], which utilizes a fixed proportion of mixed parts for auxiliary features, our approach employs bidirectional multi-step generation for multiple intermediate steps. This bolsters the robustness of V-I ReID training for samples with large discrepancies, especially at the beginning of training when the parts are inconsistent. Our strategy enables effective bridging of cross-modality gaps and allows gradual training by adjusting the proportion of information to be exchanged. We also propose a hierarchical prototype learning strategy, which is crucial to defining meaningful bridging steps.

**Our main contributions are summarized as follows.**

- (1) A BMDG method is proposed to train a backbone V-I ReID model that learns a common feature embedding by gradually reducing the gap between I and V modalities through multiple auxiliary intermediate steps.
- (2) A novel hierarchical prototype learning module is introduced to align discriminative and interchangeable part prototypes between modalities using two-level contrastive learning. These prototypes maintain semantic consistency to facilitate the exchange of V and I information, thus enabling the generation of intermediate bridging domains.
- (3) A bidirectional multi-step learning model is proposed to progressively combine prototypes from both modalities and create intermediate feature spaces at each step. It integrates attentive prototype embedding (APE), which refines these prototypes to extract attributes that are shared between modalities, thereby improving ReID accuracy. Indeed, BMDG can find better modality-shared attributes by optimizing the model on two levels: prototype alignment and bidirectional multi-step embedding modules.
- (4) Extensive experiments on the challenging SYSU-MM01 [36], RegDB [27] and LLCM [52] datasets indicate that

BMDG outperforms state-of-the-art methods for V-I person ReID. They also show that our framework can be integrated into any part-based V-I ReID method and can improve the performance of cross-modal retrieval applications.

## 2. Related Work

In V-I ReID, deep backbone models are trained with labeled V and I images captured using RGB and IR cameras. During inference, cropped images of people are matched across camera modalities. That is, query I images are matched against V gallery images, or vice versa [6, 9, 14, 15, 28, 41, 43, 59]. State-of-the-art methods can be classified into approaches for global or part-based representation, or that leverage an intermediate modality.

**(a) Global and Part-Based Representation Methods:** Representation approaches focus on training a backbone model to learn a discriminant representation that captures the essential features of each modality while exploiting their shared information [9, 23, 38, 43, 44, 47]. In [5, 28, 45, 47, 48], authors only use global features through multimodal fusion. The absence of local information limits their accuracy in more challenging cases. To improve robustness, recent methods combine global information with local part-based features [12, 18, 25, 40, 46]. For example, [25, 46] divides the spatial feature map into fixed horizontal sections and applies a weighted-part aggregation. [12, 39, 40, 58] dynamically detects regions in the spatial feature map to alleviate the misalignment of fixed horizontally divided body parts.

However, these part detectors might over-focus on specific parts, varying based on the modality or person. [18] addresses this issue using Part-Mix data augmentation to regularize model part discovery. Additionally, parts may not correspond to semantically similar regions across different images. To tackle this, a bipartite graph-based correlation model between part regions has been proposed to maximize similarities [39]. While part-based methods enhance representation ability, they often overlook shared information between similar semantic parts across different individuals or modalities, leading to inconsistencies in extracted part features. This hinders effective comparison or exchange.

Similar to the human perception system, distinguishing individuals by comparing attributes of similar body parts (e.g., head, torso, leg) learned from a diverse set of individuals, our approach seeks to capture and leverage this shared information for accurate matching. To achieve this, BMDG discovers different body parts as prototypes, disregarding the person’s identity using hierarchical contrastive learning, and then leverages their ID-discriminative attributes for each part through part-based cross-entropy loss.

**(b) Methods Based on an Intermediate Modality:** In recent years, V-I ReID methods have relied on an intermediate modality to address the significant gap between V and I modalities [2, 3, 11, 13, 20, 37, 48, 54]. Using fixed inter-

mediate modality like grayscale [11, 48] or random channels [2, 45] can improve accuracy. [20] transform V images into a new modality to reduce the gap with I images. [37] conducted a pixel-to-pixel feature fusion operation on V and I images to build the synthetic images, and [54] used a shallow auto-encoder to generate intermediate images from both modalities. While these generated images bridge the cross-modal gap, details of the image are lost. Accuracy is degraded in the context of multi-step learning. To address this, we extract body part prototypes and gradually use them to define multiple intermediate feature spaces.

**(c) Domain Generalization and Adaptation:** Data augmentation is an effective strategy to learn domain-invariant features and improve model generalization [55]. DG methods focus on manipulating the inputs to assist in learning general representations. [56] relies on domain information to create additive noise that increases the diversity of training data while preserving its semantic information. [50, 57] uses Mixup to increase data diversity by blending examples from the different training distributions.

Some domain adaptation (DA) methods allow for bridging significant shifts between source and target data distributions. Gradual and multi-step (or transitive) methods [34] define the number and location of intermediate domains. Intermediate Domain Labeler [4] also creates multiple intermediate domains, where a coarse domain discriminator sorts them such that the cycle consistency of self-training is minimized. IDM [10] mixes entire latent features to create an intermediate domain that lies on the shortest path between the target and source. In fact, by utilizing appropriate intermediate domains, the source knowledge can be better transferred to the target domain. By mixing the entire hidden features, this domain may lose the local ID-discriminative information and reduce the diversity of part features. This paper reduces the cross-modality gap by exchanging features between domains and creating intermediate virtual domains through gradual mixing during training.

## 3. Proposed Method

Our BMDG approach relies on multiple virtual intermediate domains created from V and I images to learn a common representation space for V-I person ReID. Fig. 2(a) shows the overall BMDG training architecture comprising two modules. (1) Our **part prototype alignment learning** module extracts semantically aligned and discriminative part prototypes from I and V modalities through hierarchical contrastive learning. Each prototype represents a specific part feature in a cropped person image. Exchanging aligned part prototypes allows for the gradual creation of ID-informative intermediate spaces. This module discovers and aligns part features that are distinct and informative about the person to transform the representation feature of one person between modalities robustly. (2) Our **bidirec-**

**tional multi-step learning** module creates the auxiliary intermediate feature spaces at each step by progressively mixing more learned part prototypes from each modality. This gradually reduces the modality-specific information in the final feature representation. In other words, two intermediate feature spaces are generated, at each step, by combining proportions of aligned sub-features from each modality. The model gradually learns to reduce modality-specific information and to leverage more common information from such intermediate domains by increasing the portions and making more complex training cases. BMDG trains the feature backbone by learning from easy samples with a lower modality gap to more complex ones with a higher gap.

**Preliminaries:** For the training process, let us assume a multimodal dataset for V-I ReID, represented by  $\mathcal{V} = \{x_v^j, y_v^j\}_{j=1}^{N_v}$  and  $\mathcal{I} = \{x_i^j, y_i^j\}_{j=1}^{N_i}$  which contains sets of V and I images from  $C_y$  distinct individuals, with their ID labels. V-I ReID systems seek to match images captured from V and I cameras by training a deep backbone model to encode modality-invariant person embedding, denoted by  $\mathbf{f}_v$  and  $\mathbf{f}_i$ . Given V and I images, the objective is to:

$$\max\{S(\mathbf{f}_v^n, \mathbf{f}_i^l) \cdot o\} \text{ where } o = 1 \text{ if } y_v^n = y_i^l \text{ else } o = -1, \quad (1)$$

where  $\mathbf{f}_v^n, \mathbf{f}_i^l \in \mathbb{R}^d$ ,  $d$  is the dimension of the representation space, and  $S(\cdot, \cdot)$  is similarity matching function.

### 3.1. Prototype Learning Module

To align persons' sub-features between different images in I and V, this module is used to discover different body-part attributes from diverse regions and learn to align these prototypes by making them complementary, interchangeable, and discriminative, as discussed in Section 1.

$$\mathcal{H}: \mathbb{R}^{W \times H \times 3} \rightarrow \mathbb{R}^{K \times d} \quad (2)$$

$$x \sim X \mapsto \mathcal{H}(x) \sim P \quad (3)$$

**(a) Prototype Discovery (PD).** This module is proposed to extract the prototypes and discover several discriminative regions on the backbone feature maps. At first, the PD predicts a pixel-wise masking score (probability)  $M \in \mathbb{R}^{H \cdot W \times K}$  for each prototype from the features map  $\mathbf{F}_m \in \mathbb{R}^{H \cdot W \times d}$  where  $H, W, d$  are the feature sizes, and  $K$  is the total number of prototypes.  $m$  is modality index which indicates  $v$  as visible or  $i$  as infrared domain. Then, it computes each prototype vector  $\mathbf{p}_m^k$  by weighted aggregating of features in all pixels as:

$$\mathbf{p}_m^k = \frac{M_m^k}{|M_m^k|} \sum_{u \in U} [\mathbf{F}_m]_u, \quad |M_m^k| = \sum_{u \in U} [M_m^k]_u \quad (4)$$

where for each pixel  $u \in U = \{1, \dots, H \cdot W\}$ , we have  $\sum_{k=1}^K [M_m^k]_u = 1$ . This module (illustrated in Fig. 2(b)) uses a shallow U-Net [29] architecture to create a region mask for each prototype. In the down-sample block, a convolution receptive field is applied on masks to make

them lose locality by leveraging their neighbor pixels. At the same time, the up-sample propagates these high-level, coarsely localized masks into each original size. Skip connections are used to re-inject the local details lost in the down-sample phase.

**(b) Hierarchical Contrastive Learning:** To ensure the part prototypes represent the diverse attributes of each person, they must be non-redundant, and be independent of the others. This independence can be formulated that the mutual information ( $MI$ ) between the distribution of random variables describing parts as  $P^k$  should be zero as:

$$MI(P^k; P^q) = 0 \quad \forall k, q \in \{1, \dots, K\}, k \neq q, \quad (5)$$

where  $P^k$  is the distribution of the random variable of part  $k$ . Since the estimation of mutual information is complex and time-consuming,  $MI(P^k, P^q)$  is minimized by directly reducing the cosine similarity between the extracted features of different parts as:

$$\min S\{\{\mathbf{p}^k, \mathbf{p}^q\}\} \quad \forall k, q \in \{1, \dots, K\}, k \neq q, \quad (6)$$

where  $\mathbf{p}^k$  and  $\mathbf{p}^q$  are features for part  $k$  and  $q$ , respectively. To ensure that the prototype features are semantically consistent around body parts that can be aligned in different samples, we maximize the similarity of  $\mathbf{p}^k$  and all other prototypes with the same index in the training batch as  $\hat{\mathbf{p}}^k$ .

To select samples in each batch with a positive prototype index, there are two options: the same parts belonging to the same or different persons (inter-person) as  $\Psi^k$ , and the same parts belonging to the same person with identity  $y$  (intra-person) as  $\Gamma_y^k$  (see Fig. 2.b). Since the goal is to extract dissimilar features in the representation space for two different persons, pulling apart the same prototypes from these two persons may disrupt this objective. Therefore, two contrastive losses as  $\mathcal{L}_{lc}$  and  $\mathcal{L}_{hc}$  are employed to provide inter-person information at a low-level and intra-person at a high-level:

$$\mathcal{L}_{lc} = - \sum_{k=1}^K \sum_{\hat{\mathbf{p}}' \in \Psi^k} \log \frac{e^{\mathbf{p}^k \cdot \hat{\mathbf{p}}' / \tau}}{e^{\mathbf{p}^k \cdot \hat{\mathbf{p}}' / \tau} + \sum_{q \neq k} e^{\mathbf{p}^k \cdot \mathbf{p}^q / \tau}}, \quad (7)$$

$$\mathcal{L}_{hc} = - \sum_{y=1}^{C_y} \sum_{k=1}^K \sum_{\hat{\mathbf{p}} \in \Gamma_y^k} \log \frac{e^{\mathbf{p}^k \cdot \hat{\mathbf{p}} / \tau}}{e^{\mathbf{p}^k \cdot \hat{\mathbf{p}} / \tau} + \sum_{q \neq k} e^{\mathbf{p}^k \cdot \mathbf{p}^q / \tau}}, \quad (8)$$

where  $\mathbf{p}_m^{k'}$  are low-level part prototype features vector computed same as Eq. (4) but extracted from lower-level layers of the feature backbone. Unlike contrastive learning introduced in [8, 22] that are used only for final features, the HCL optimizes at two levels as low-level and high-level features to force the model to cluster inputs based on parts and then cluster inside identities, as illustrated in Fig. 2(b). For example,  $\mathcal{L}_{lc}$  encourages the low-level prototype of legs for two persons to be similar and differ from other prototypes such as the torso. While  $\mathcal{L}_{hc}$  encourages the high-level

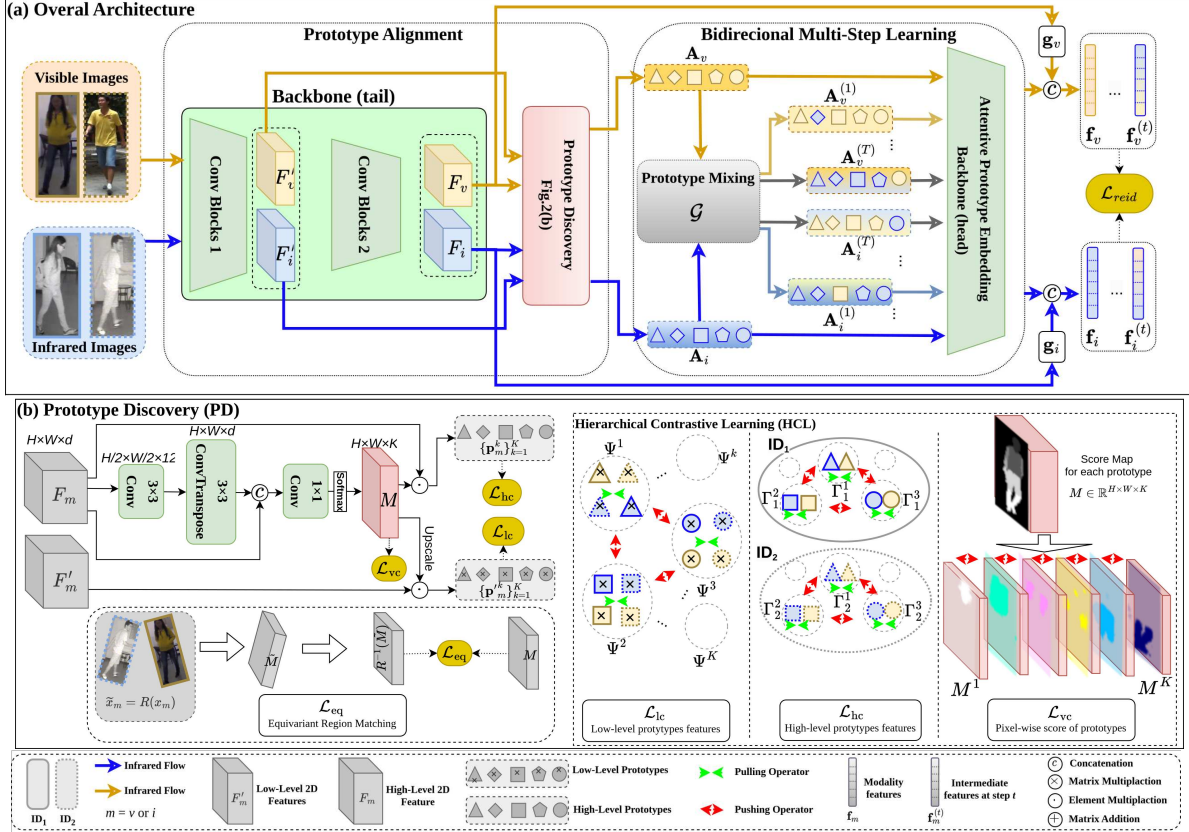


Figure 2. (a) Overall BMDG training architecture comprises two parts. The prototype alignment module (left) extracts body part prototype representations from V and I images. The bidirectional multi-step learning module (right) extracts discriminant features using multiple intermediate domains created by mixing prototype information. (b) Prototype discovery (PD) architecture mines prototypes from spatial features, and Hierarchical contrastive learning (HCL) encourages the prototypes to focus on similar semantics for all individuals without losing ID-discriminative information.

legs prototypes only differ from other prototypes in that person. Also, minimizing the distance between features of pixels belonging to the same part occurrence encourages the model to extract more similar features for each prototype:

$$\mathcal{L}_c = \sum_{k=1}^K \sum_{u \in U} \|M_m^k - \mathbf{F}_m\|_u. \quad (9)$$

Describing part occurrence to a single feature vector enables us to exchange these features for the corresponding prototype in different modalities and creates the intermediate step. Apart from making part features independent of each other by feature contrastive losses, for pushing prototypes to focus on different semantic body parts in different image locations, the visual contrastive is used as:

$$\mathcal{L}_{vc} = \sum_{k=1}^K \sum_{q=k+1}^K \sum_{u \in U} \|M_m^k - M_m^q\|_u, \quad (10)$$

to extract parts from diverse regions of backbone features with the minimum intersection. Also, for equivariant matching of regions, a random rigid transformation ( $R$ ) is applied on the input images ( $x_m$ ) to create the transformed

images ( $\tilde{x}_m$ ). Both are then processed by the model, and the transformation on the masking score ( $\tilde{M}_m$ ) is inverted from the transformed images. The objective is to minimize the distance between the transformed and original maps as:

$$\mathcal{L}_{eq} = \sum_{k=1}^K \|M_m^k - R^{-1}(\tilde{M}_m^k)\|. \quad (11)$$

**(c) Part Discrimination.** The information encoded by prototypes must describe the object in the foreground. That is, the mutual information between the joint probability  $\{P^k\}_{k=1}^K$  and identity  $Y$  should be maximized:

$$MI(P^1, \dots, P^K; Y). \quad (12)$$

In the supplementary material, we proved that for maximizing Eq. (12), we can minimize cross-entropy loss between each prototypes features and identities. In fact, by doing this, the model makes the part prototype regions lie on the object and have enough information about the foreground. In other words, each part prototype should be able to recover the identity of the person in the images. So part ID-discriminative loss is defined as:

$$\mathcal{L}_p = \frac{-1}{K} \sum_{k=1}^K \sum_{c=1}^{C_y} y_c \log(W_c^k(\mathbf{p}_m^k)), \quad (13)$$

where  $W^k$  is linear layer predicting the probability of identity  $y_c$  from  $\mathbf{p}_m^k$ . To ensure prototype diversity and discrimination, we do not share parameters of  $W^k$  and  $W^q$  and randomly drop out a proportion of them during training, preventing any single part from dominating.

### 3.2. Bidirectional Multi-step Learning

After identifying prototypes, multiple auxiliary intermediate domains are created, and features are extracted using the Attentive Prototype Embedding (APE) module.

**(a) Attentive Prototype Embedding.** Instead of directly concatenating learned prototypes as the final feature, we process all other prototypes in input images. The motivation of this design is that each prototype may discriminate the part-attributes shared for all individuals rather than ID attributes. The Attentive Prototype Embedding module (which is detailed in the supplementary materials),  $\mathcal{F}$ , leverages relevant person information between prototypes by applying an attention mechanism to achieve person-aware final features. To aggregate the ID-discriminative features of prototypes, the APE module first uses a fully connected layer to score and emphasize important channels in their feature map. It weights each prototype feature by computing the similarities between them.

$$\mathcal{F}(\mathbf{A}_m) = \mathbf{W}_{\text{mlp}}(\mathbf{C}_m), \text{ where } \mathbf{C}_m = \mathbf{W}_v(\mathbf{A}_m) \otimes \mathbf{B}_m, \\ \mathbf{B}_m = \sigma(\mathbf{W}_q(\mathbf{A}_m) \otimes \mathbf{W}_k(\mathbf{A}_m)), \quad (14)$$

and

$$\mathbf{A}_m = [\mathbf{p}_m^1; \mathbf{p}_m^2; \dots; \mathbf{p}_m^K], \quad (15)$$

where  $\mathbf{A}_m \in \mathbb{R}^{K \times d}$ ,  $\otimes$  is matrix multiplication and  $\sigma$  is the Sigmoid function.  $\mathbf{W}_{\text{mlp}}$ ,  $\mathbf{W}_v$ ,  $\mathbf{W}_q$  and  $\mathbf{W}_k$  are linear layer. Also, to increase the discriminative ability of the final feature embedding, the global features,  $\mathbf{g}_m$ , extracted by the backbone head are concatenated to the output of APE:

$$\mathbf{f}_m = [\mathcal{F}(\mathbf{A}_m); \mathbf{g}_m], \quad (16)$$

where  $\mathbf{g}_m$  is denoted by  $\frac{1}{H \cdot W} \sum_{u \in U} [\mathbf{F}_m]_u$ .

Our APE has two advantages: (1) it allows adjusting modality-agnostic attention between prototypes regardless of the modality by applying cross-modality prototype features, and (2) it improves the representation ability of final embedding by emphasizing most discriminative prototypes.

**(b) Bidirectional Multi-Step Learning.** To deal with the significant shift between modalities in the feature space, the model is trained via multiple intermediate steps that gradually bridge this domain gap. Using intermediates with less domain shift, the model gradually learns to leverage cross-modality discriminating clues at each step [1, 4, 33, 51]. Initially, the discrepancies between modalities are small, letting the model learn from the easier samples first, then converge on more complex cases with larger shifts. One solution to achieve these intermediates is to start from one

modality as the source and transform gradually to the other as the target. However, this makes the model forget the learned knowledge of the source and be biased on the target.

To address this issue, modalities are transformed bidirectionally to create an intermediate domain at each step. By gradually transforming domains to each other, the domain gap vanishes over multiple steps. Each intermediate auxiliary step provides discriminative information about the person across both modalities, enabling the training process to transform inputs from one to the other. By creating these intermediate domains, our model aims to neither lose nor duplicate information from the main domains. For gradual transformation in each step, we exchange the features of the same prototypes from both modalities to create a mixed and virtual representation space with less domain discrepancies. These prototype-mixed intermediate feature spaces are used alongside V and I prototypes to provide APE along with the ability to extract common features from the modalities. For step  $t$ , which  $t \in \{1, \dots, T\}$  and  $T$  is the number of intermediate steps, the intermediate features for the same person are generated using a mixing function  $\mathcal{G}(\cdot, \cdot, \cdot)$  as:

$$\mathbf{A}_m^{(t)} = \mathcal{G}(\mathbf{A}_m, \mathbf{A}_{\tilde{m}}, t), \quad (17)$$

where  $\tilde{m} \neq m$  which mixes the prototypes from two modalities. For example, we use a simple but yet effective random prototype mixing strategy for  $\mathcal{G}(\cdot, \cdot, \cdot)$ :

$$\mathcal{G}(\mathbf{A}_m, \mathbf{A}_{\tilde{m}}, t) = [\mathbf{p}_{r(m, \tilde{m}, t)}^1; \dots; \mathbf{p}_{r(m, \tilde{m}, t)}^K], \quad (18)$$

where  $r(\cdot, \cdot, \cdot)$  is a weighted random selector:

$$r(m, \tilde{m}, t) = \begin{cases} m & t/T \leq \mathcal{U}(0, 1) \\ \tilde{m} & \text{else} \end{cases} \quad (19)$$

where  $\mathcal{U}(0, 1)$  is a uniform random generator. For intermediate step ( $t < T$ ), we have  $\mathbf{f}_m^{(t)} = [\mathcal{F}(\mathbf{A}_m^{(t)}); \mathbf{g}_m]$  and in the last step  $\mathbf{f}_m^{(T)} = [\mathcal{F}(\mathbf{A}_m^{(T)}); \mathbf{g}_{\tilde{m}}]$  is used. The module  $\mathcal{F}$  learns to gradually reduce the discrepancies in representation space by applying metric learning objectives bidirectionally between  $\mathbf{f}_i$  and  $\mathbf{f}_i^{(t)}$  and between  $\mathbf{f}_v$  and  $\mathbf{f}_v^{(t)}$  as:

$$\mathcal{L}_{\text{bcc}} = \mathcal{L}_{\text{cc}}(\mathbf{f}_v, \mathbf{f}_v^{(t)}) + \mathcal{L}_{\text{cc}}(\mathbf{f}_i, \mathbf{f}_i^{(t)}) \\ \mathcal{L}_{\text{bce}} = \mathcal{L}_{\text{ce}}(\mathbf{f}_v) + \mathcal{L}_{\text{ce}}(\mathbf{f}_v^{(t)}) + \mathcal{L}_{\text{ce}}(\mathbf{f}_i) + \mathcal{L}_{\text{ce}}(\mathbf{f}_i^{(t)}) \quad (20) \\ \mathcal{L}_{\text{re}} = \mathcal{L}_{\text{bce}} + \mathcal{L}_{\text{bcc}},$$

that  $\mathcal{L}_{\text{cc}}$ ,  $\mathcal{L}_{\text{ce}}$  are center cluster and cross-entropy losses [40]. Since  $\mathbf{f}_i^{(t)}$  begins from  $\mathbf{f}_i$  at first steps and approaches to  $\mathbf{f}_v$  at ending steps, the model is gradually trained to extract more robust modality invariant features by seeing samples with low modality gap to the harder ones with higher gap [33].

### 3.3. Training and Inference

In each step, the model tries to represent the same feature space for two input domains, V/I and the intermediate ones, using the overall loss expressed as:

Family			SYSU-MM01						RegDB					
			All Search			Indoor Search			Visible → Infrared			Infrared → Visible		
Method		Venue	R1	R10	mAP	R1	R10	mAP	R1	R10	mAP	R1	R10	mAP
Global	AGW [47]	TPAMI'20	47.50	-	47.65	54.17	-	62.97	70.05	-	50.19	70.49	87.21	65.90
	CAJ [45]	ICCV'21	69.88	-	66.89	76.26	97.88	80.37	85.03	95.49	79.14	84.75	95.33	77.82
	RAPV-T [49]	ESA'23	63.97	95.30	62.33	69.00	97.39	75.41	86.81	95.81	81.02	86.60	96.14	80.52
	G2DA [31]	PR'23	63.94	93.34	60.73	71.06	97.31	76.01	-	-	-	-	-	-
Part-based	DDAG [46]	ECCV'20	54.74	90.39	53.02	61.02	94.06	67.98	69.34	86.19	63.46	68.06	85.15	90.31
	SSFT [25]	CVPR'20	63.4	91.2	62.0	70.50	94.90	72.60	71.0	-	71.7	-	-	-
	MPANet [40]	CVPR'22	70.58	96.21	68.24	76.74	98.21	80.95	83.70	-	80.9	82.8	-	80.7
	SAAI [12] <sup>a</sup>	ICCV'23	75.03	-	71.69	-	-	-	-	-	-	-	-	-
	SAAI [12] <sup>b</sup>	ICCV'23	75.90	-	<u>77.03</u>	82.20	-	80.01	91.07	-	<u>91.45</u>	92.09	-	<u>92.01</u>
	CAL [39]	ICCV'23	74.66	96.47	71.73	79.69	98.93	83.68	<u>94.51</u>	<b>99.70</b>	88.67	<u>93.64</u>	<b>99.46</b>	87.61
	PartMix [18]	CVPR'23	77.78	-	74.62	81.52	-	84.83	84.93	-	82.52	85.66	-	82.27
Intermediate	SMCL [35]	ICCV'21	67.39	92.84	61.78	68.84	96.55	75.56	83.93	-	79.83	83.05	-	78.57
	MMN [54]	ICM'21	70.60	96.20	66.90	76.20	<b>99.30</b>	79.60	91.60	97.70	84.10	87.50	96.00	80.50
	RPIG [2]	ECCVw'22	71.08	96.42	67.56	82.35	98.30	82.73	87.95	98.3	82.73	86.80	96.02	81.26
	FTMI [30]	MVA'23	60.5	90.5	57.3	-	-	-	79.00	91.10	73.60	78.8	91.3	73.7
	G2DA [31]	PR'23	63.94	93.34	60.73	71.06	97.31	76.01	-	-	-	-	-	-
	SEFL [13]	CVPR'23	75.18	96.87	70.12	78.40	97.46	81.20	91.07	-	85.23	92.18	-	86.59
	BMDG (ours) <sup>a</sup>	-	76.15	<u>97.42</u>	73.21	<u>83.53</u>	98.02	<u>84.92</u>	93.92	98.11	89.18	<u>94.08</u>	97.0	88.67
	BMDG (ours) <sup>b</sup>	-	<b>78.08</b>	<b>97.90</b>	<b>78.22</b>	<b>83.59</b>	<u>98.96</u>	<b>86.35</b>	<b>94.76</b>	<u>98.91</u>	<b>92.21</b>	<b>94.56</b>	<u>98.31</u>	<b>93.07</b>

<sup>a</sup> without AIM <sup>b</sup> with AIM

Table 1. Accuracy of the proposed BMDG and state-of-the-art methods on the SYSU-MM01 (single-shot setting) and RegDB datasets. All numbers are percent. Results for methods were obtained from the original papers. AIM [12] means re-ranking method.

$$\mathcal{L} = \mathcal{L}_{re} + \lambda_f(\mathcal{L}_{lc} + \mathcal{L}_{hc}) + \lambda_v\mathcal{L}_{vc} + \lambda_c\mathcal{L}_c + \lambda_i\mathcal{L}_p + \lambda_e\mathcal{L}_{eq} \quad (21)$$

where  $\lambda$  are hyper-parameters for weighting losses. During inference, the tail and head backbones are combined to extract the prototypes and global features as  $\mathbf{A}_i$ ,  $\mathbf{A}_v$ ,  $\mathbf{g}_i$  and  $\mathbf{g}_v$  from given input image  $x_i$  and  $x_v$ . Using Eq. (16), the  $\mathbf{f}_i$  and  $\mathbf{f}_v$  are computed by the APE. The matching score for a query input image is computed by cosine similarity.

## 4. Results and Discussion

In this section, we compare the BMDG with SOTA V-I ReID methods, including global, part-based, and intermediate approaches. Extensive ablation studies are conducted to show the effectiveness of the proposed bi-directional learning, the impact of the number of parts-prototypes, and intermediate steps during learning. The supplementary materials provide detailed information on our experimental methodology, including the implementation of baseline models, performance measures, and descriptions of the SYSU-MM01 [36], RegDB [27], and LLCM [52] datasets.

Family		LLCM			
		V → I		I → V	
Method	Venue	R1	mAP	R1	mAP
DDAG [46]	ECCV'20	40.3	48.4	48.0	52.3
CAJ [45]	ICCV'21	56.5	59.8	48.8	56.6
RPIG [2]	ECCVw'22	57.8	61.1	50.5	58.2
DEEN [53]	CVPR'23	62.5	65.8	54.9	62.9
BMDG (ours)	-	<b>63.4</b>	<b>66.3</b>	<b>56.4</b>	<b>63.5</b>

Table 2. Accuracy of the proposed BMDG and state-of-the-art methods on the Large LLCM Dataset. All numbers are percent.

## 4.1. Comparison with State-of-Art Methods:

Table 1 compares the accuracy of BMDG with state-of-the-art V-I ReID approaches. Our experiments on the SYSU-MM01 and RegDB datasets show that BMDG outperforms the SOTA methods in the majority of situations. Compared to [18], BMDG has lower R1 accuracy (-1%) in the All Search settings on the SYSU-MM01 dataset. However, it achieves higher performance in Indoor settings (+2%) and on the RegDB dataset (+10%). These results can be attributed to BMDG's ability to effectively capture more ID-related knowledge across modalities. This is achieved by learning discriminative part-prototypes and gradually reducing modality-specific information in the extracted final features. Additionally, the prototype alignment module creates gradual and bidirectional intermediate spaces without sacrificing discriminative ability, allowing BMDG to surpass intermediate methods. Moreover, BMDG can be easily integrated into different part-based V-I ReID models, enhancing generality through gradual training. To show its effectiveness, we integrated our approach into state-of-the-art part-based or prototype-based baseline models [12, 40, 46]. We executed the original code published by authors using the hyper-parameters and other configurations they provided with and without BMDG. Table 3 shows that BMDG improves performance for mAP and rank-1 accuracy for all methods by an average of 2.02% and 1.54%, respectively. Similar performance improvements are observed on the large and complex LLCM dataset (see Table 2). Since LLCM was introduced recently, few papers reported their results, so we executed other approaches with published code on the LLCM dataset.

Method	R1 (%)	mAP (%)
DDAG [46]	53.62	52.71
DDAG with BMDG	<b>55.36 (+1.74↑)</b>	<b>54.05 (+1.34↑)</b>
MPANet [40]	66.24	62.89
MPANet with BMDG	<b>68.74 (+2.50↑)</b>	<b>64.25 (+1.36↑)</b>
SAAI [12]	71.87	68.16
SAAI with BMDG	<b>73.69 (+1.82↑)</b>	<b>70.08 (+1.92↑)</b>

Table 3. Accuracy of part-based ReID methods with BMDG on the SYSU-MM01, under single-shot setting. Results were obtained by executing the author’s code on our servers.

## 4.2. Ablation Studies:

**(a) Step size and number of part prototypes.** We evaluate the best option for the number of parts and steps. Table 4 (top) shows that using the BMDG approach increases the R1 accuracy when increasing the number of steps  $T$  for a specific number of prototypes  $K$ . Also, it increases performance in multi-step with  $K$  parts, letting the model learn from the lower gap sample and converge on harder cases.

Another option for each step, instead of exchanging some parts, is to use the Mixup [50] or IDM [10] strategy for all parts from two modalities with normalized weight:  $\mathbf{A}_m^{(t)} = \alpha^{(t)} \mathbf{A}_m + (1 - \alpha^{(t)}) \mathbf{A}_{m'}$ , where  $\alpha$  starts from 0 to 1 w.r.t. step  $t$  and number of steps. Results in Table 4 (bottom) show that using multi-step increases performance marginally since it does not utilize the ability of prototypes to create meaningful intermediate steps. This is explained by the fact that prototypes contain some modality-specific information, which is crucial for transforming domains in intermediate steps and may be lost in the averaging process.

**(c) Loss functions.** Contrastive objectives between parts ( $\mathcal{L}_{lc}$  and  $\mathcal{L}_{hc}$ ) are essential for the model to detect similar semantics between objects, thereby creating an intermediate step by swiping corresponding parts. Table 5 shows this necessity when used in conjunction with  $\mathcal{L}_c$ . Using such regularization in the multi-step approach increases mAP by

$T$	Number of part prototypes ( $K$ )					
	3	4	5	6	7	10
<b>Prototype exchanging (Our)</b>						
0	68.25	69.24	69.40	70.27	70.03	68.12
1	69.97	70.81	72.07	71.97	71.31	69.28
2	71.20	72.35	73.98	73.61	72.45	71.25
3	73.32	73.94	74.11	74.98	73.22	71.67
4	-	74.08	74.15	<b>75.43</b>	73.51	71.99
6	-	-	-	75.37	73.52	72.15
10	-	-	-	-	-	72.07
<b>Entire mixup (IDM [10])</b>						
0	68.25	69.24	69.4	70.27	70.03	68.12
1	68.53	69.19	69.56	70.53	70.3	68.41
4	68.70	70.03	69.75	70.84	70.77	68.53
6	69.13	70.81	70.1	<b>71.16</b>	70.29	68.90
10	69.21	70.15	69.52	70.95	70.90	68.13

Table 4. R1 accuracy of BMDG using (a) our prototype mixing and (b) IDM [10] for different numbers of part prototypes ( $K$ ) and intermediate steps ( $T$ ). mAP is reported in suppl. materials.

Losses						R1 (%)		mAP (%)	
$\mathcal{L}_p$	$\mathcal{L}_{lc}$	$\mathcal{L}_{hc}$	$\mathcal{L}_{vc}$	$\mathcal{L}_c$	$\mathcal{L}_{eq}$	$T=0$	$T=K$	$T=0$	$T=K$
baseline						51.09	-	49.68	-
	✓	✓				62.19	65.28	58.91	60.73
	✓	✓	✓			66.35	69.50	63.01	66.48
				✓		66.67	67.31	61.04	61.68
✓					✓	59.78	60.66	55.41	56.05
✓	✓	✓	✓			68.30	69.04	66.87	66.99
	✓	✓	✓	✓		69.68	71.20	67.08	68.13
✓	✓	✓	✓	✓		69.90	72.67	68.11	69.75
✓	✓	✓	✓	✓	✓	<b>70.27</b>	<b>75.37</b>	<b>68.45</b>	<b>72.71</b>

Table 5. Impact on the accuracy of using different BMDG losses, using SYSU-MM01 for single-step ( $T=0$ ) and multi-step ( $T=K$ ) settings. The baseline  $\mathcal{L}_{re}$  loss is used in all cases.

3% and R1 by 4%. Also, results of part separation loss,  $\mathcal{L}_{vc}$ , indicate that using this loss pushes the parts to be scattered on the body regions and gives the part features a more robust representation. Additionally, it enables the intermediate steps to generalize the feature space better by increasing the performance over the single-step with multi-step setting. The part identity loss  $\mathcal{L}_p$  lets the model extract ID-related prototype features that contain discriminative information about the object and generate an id-informative intermediate step. Without this objective, parts features are likely to focus on background information. As shown in the two last rows, performance increases when using this objective.

**(d) Bidirectional domain generalization.** To show the impact of the BMDG approach, we conduct experiments under the following settings: (1) single-step, (2) one-directional multi-step (from  $V \rightarrow I$  and  $I \rightarrow V$ ), and (3) bidirectional for creating the intermediate series. The results are shown in Table 6. Although using multiple intermediate steps from one modality to another improves the model’s accuracy by allowing it to gradually learn and utilize common discriminative features, this process can lead to losing information about the source modality. The bi-directional approach addresses these issues and achieves the best performance, as it is not biased towards any specific modality.

Settings	Number of part prototypes ( $K$ )							
	4		5		6		7	
	R1	mAP	R1	mAP	R1	mAP	R1	mAP
Single step	69.2	65.9	69.4	66.1	70.2	66.3	70.4	66.5
One ( $V \rightarrow I$ )	71.0	66.5	71.5	67.1	72.3	67.5	71.6	66.8
One ( $I \rightarrow V$ )	70.8	66.3	71.2	67.0	72.6	67.6	71.0	66.7
Bidirectional	<b>74.0</b>	<b>71.8</b>	<b>74.1</b>	<b>71.9</b>	<b>75.4</b>	<b>72.8</b>	<b>73.5</b>	<b>70.2</b>

Table 6. Impact on rank-1 and mAP accuracy using different settings for multi-step domain generalization. All numbers are percent.  $V \rightarrow I$  means that the intermediates are created only from  $V$ , by replacing prototypes from  $I$  features.

## 5. Conclusion

This paper introduces the BMDG framework for V-I ReID. It uses a novel prototype learning approach to learn the most discriminative and complementary set of prototypes from both modalities. These prototypes generate multiple intermediate domains between modalities by progres-

sively mixing prototypes from each modality, reducing the domain gap. The effectiveness of BMDG is shown experimentally on several datasets, where it outperforms state-of-the-art methods for V-I person ReID. Results also highlight the impact of considering multiple intermediate steps and bidirectional training to improve accuracy. Moreover, BMDG can be integrated into other part-based V-I ReID methods, significantly improving their accuracy. A limitation of BMDG is its dependency on discriminant body parts. **Acknowledgements.** This research was supported by the Natural Sciences and Eng. Research Council of Canada.

## References

- [1] Samira Abnar, Rianne van den Berg, Golnaz Ghiasi, Mostafa Dehghani, Nal Kalchbrenner, and Hanie Sedghi. Gradual domain adaptation in the wild: When intermediate distributions are absent. *arXiv preprint arXiv:2106.06080*, 2021. [6](#)
- [2] Mahdi Alehdaghi, Arthur Josi, Rafael MO Cruz, and Eric Granger. Visible-infrared person re-identification using privileged intermediate information. In *ECCVws*, pages 720–737. Springer, 2022. [1](#), [3](#), [7](#)
- [3] Mahdi Alehdaghi, Arthur Josi, Pourya Shamsolmoali, Rafael MO Cruz, and Eric Granger. Adaptive generation of privileged intermediate information for visible-infrared person re-identification. *arXiv preprint arXiv:2307.03240*, 2023. [3](#)
- [4] Hong-You Chen and Wei-Lun Chao. Gradual domain adaptation without indexed intermediate domains. *Advances in Neural Information Processing Systems*, 34:8201–8214, 2021. [3](#), [6](#)
- [5] Kunfeng Chen, Zhisong Pan, Jiabao Wang, Shanshan Jiao, Zhicheng Zeng, and Zhuang Miao. Joint feature learning network for visible-infrared person re-identification. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 652–663. Springer, 2020. [3](#)
- [6] Yehansen Chen, Lin Wan, Zhihang Li, Qianyan Jing, and Zongyuan Sun. Neural feature search for rgb-infrared person re-identification. In *CVPR*, pages 587–597, 2021. [3](#)
- [7] Seokeon Choi, Sumin Lee, Youngeun Kim, Taekyung Kim, and Changick Kim. Hi-cmd: hierarchical cross-modality disentanglement for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10257–10266, 2020. [1](#)
- [8] Subhabrata Choudhury, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Unsupervised part discovery from contrastive reconstruction. *Advances in Neural Information Processing Systems*, 34:28104–28118, 2021. [4](#)
- [9] 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. [3](#)
- [10] Yongxing Dai, Jun Liu, Yifan Sun, Zekun Tong, Chi Zhang, and Ling-Yu Duan. Idm: An intermediate domain module for domain adaptive person re-id. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11864–11874, 2021. [3](#), [8](#)
- [11] Xing Fan, Haiyan Luo, Chi Zhang, and Wei Jiang. Cross-spectrum dual-subspace pairing for rgb-infrared cross-modality person re-identification. *ArXiv*, abs/2003.00213, 2020. [3](#)
- [12] Xingye Fang, Yang Yang, and Ying Fu. Visible-infrared person re-identification via semantic alignment and affinity inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11270–11279, 2023. [2](#), [3](#), [7](#), [8](#)
- [13] Jiawei Feng, Ancong Wu, and Wei-Shi Zheng. Shape-erased feature learning for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22752–22761, 2023. [3](#), [7](#)
- [14] Chaoyou Fu, Yibo Hu, Xiang Wu, Hailin Shi, Tao Mei, and Ran He. Cm-nas: Cross-modality neural architecture search for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11823–11832, 2021. [3](#)
- [15] Yi Hao, Jie Li, Nannan Wang, and Xinbo Gao. Modality adversarial neural network for visible-thermal person re-identification. *Pattern Recognition*, 107:107533, 2020. [3](#)
- [16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. [5](#)
- [17] Zhipeng Huang, Jiawei Liu, Liang Li, Kecheng Zheng, and Zheng-Jun Zha. Modality-adaptive mixup and invariant decomposition for rgb-infrared person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1034–1042, 2022. [1](#)
- [18] Minsu Kim, Seungryong Kim, Jungin Park, Seongheon Park, and Kwanghoon Sohn. Partmix: Regularization strategy to learn part discovery for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18621–18632, 2023. [2](#), [3](#), [7](#)
- [19] Vladimir V Kniaz, Vladimir A Knyaz, Jirí Hladuvka, Walter G Kropatsch, and Vladimir Mizginov. Thermalgan: Multimodal color-to-thermal image translation for person re-identification in multispectral dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018. [1](#)
- [20] Diangang Li, Xing Wei, Xiaopeng Hong, and Yihong Gong. Infrared-visible cross-modal person re-identification with an x modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4610–4617, 2020. [1](#), [3](#)
- [21] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. Self-correction for human parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. [8](#), [9](#)
- [22] Yulin Li, Jianfeng He, Tianzhu Zhang, Xiang Liu, Yongdong Zhang, and Feng Wu. Diverse part discovery: Occluded person re-identification with part-aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2898–2907, 2021. [4](#)

- [23] Haijun Liu and Jian Cheng. Enhancing the discriminative feature learning for visible-thermal cross-modality person re-identification. *CoRR*, abs/1907.09659, 2019. 3
- [24] Haojie Liu, Shun Ma, Daoxun Xia, and Shaozi Li. Sfanet: A spectrum-aware feature augmentation network for visible-infrared person re-identification. *arXiv preprint arXiv:2102.12137*, 2021. 2
- [25] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13379–13389, 2020. 1, 3, 7
- [26] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. 6, 7
- [27] 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. 2, 7, 5
- [28] Hyunjong Park, Sanghoon Lee, Junghyup Lee, and Bumsub Ham. Learning by aligning: Visible-infrared person re-identification using cross-modal correspondences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12046–12055, 2021. 1, 3
- [29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015 Conference Proceedings, 2022*. 4
- [30] Jia Sun, Yanfeng Li, Houjin Chen, Yahui Peng, and Jinlei Zhu. Visible-infrared person re-identification model based on feature consistency and modal indistinguishability. *Machine Vision and Applications*, 34(1):14, 2023. 7
- [31] Lin Wan, Zongyuan Sun, Qianyan Jing, Yehansen Chen, Lijing Lu, and Zhihang Li. G2da: Geometry-guided dual-alignment learning for rgb-infrared person re-identification. *Pattern Recognition*, 135:109150, 2023. 7
- [32] 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 *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 1
- [33] Haoxiang Wang, Bo Li, and Han Zhao. Understanding gradual domain adaptation: Improved analysis, optimal path and beyond. In *International Conference on Machine Learning*, pages 22784–22801. PMLR, 2022. 6
- [34] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018. 3
- [35] Ziyu Wei, Xi Yang, Nannan Wang, and Xinbo Gao. Syncretic modality collaborative learning for visible infrared person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 225–234, 2021. 7
- [36] 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. 2, 7, 5
- [37] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. Rgb-infrared cross-modality person re-identification. In *Proceedings of the IEEE international conference on computer vision*, pages 5380–5389, 2017. 1, 3
- [38] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. Rgb-infrared cross-modality person re-identification. In *Proceedings of the IEEE international conference on computer vision*, pages 5380–5389, 2017. 3
- [39] Jianbing Wu, Hong Liu, Yuxin Su, Wei Shi, and Hao Tang. Learning concordant attention via target-aware alignment for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11122–11131, 2023. 3, 7
- [40] Qiong Wu, Pingyang Dai, Jie Chen, Chia-Wen Lin, Yongjian Wu, Feiyue Huang, Bineng Zhong, and Rongrong Ji. Discover cross-modality nuances for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4330–4339, 2021. 1, 3, 6, 7, 8
- [41] Xiaohui Xu, Song Wu, Shan Liu, and Guoqiang Xiao. Cross-modal based person re-identification via channel exchange and adversarial learning. In *International Conference on Neural Information Processing*, pages 500–511. Springer, 2021. 3
- [42] Yang Yang, Tianzhu Zhang, Jian Cheng, Zengguang Hou, Prayag Tiwari, Hari Mohan Pandey, et al. Cross-modality paired-images generation and augmentation for rgb-infrared person re-identification. *Neural Networks*, 128:294–304, 2020. 1
- [43] Mang Ye, Xiangyuan Lan, Jiawei Li, and Pong Yuen. Hierarchical discriminative learning for visible thermal person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 3
- [44] M. Ye, X. Lan, Z. Wang, and P. C. Yuen. Bi-directional center-constrained top-ranking for visible thermal person re-identification. *IEEE Transactions on Information Forensics and Security*, 15:407–419, 2020. 1, 3
- [45] Mang Ye, Weijian Ruan, Bo Du, and Mike Zheng Shou. Channel augmented joint learning for visible-infrared recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13567–13576, 2021. 3, 7
- [46] Mang Ye, Jianbing Shen, David J. Crandall, Ling Shao, and Jiebo Luo. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 229–247, Cham, 2020. Springer International Publishing. 1, 2, 3, 7, 8
- [47] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C. H. Hoi. Deep learning for person re-identification: A survey and outlook. *arXiv preprint arXiv:2001.04193*, 2020. 1, 3, 7, 6, 8
- [48] 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. 1, 3

- [49] Xuanli Zeng, Jun Long, Shengwei Tian, and Guangli Xiao. Random area pixel variation and random area transform for visible-infrared cross-modal pedestrian re-identification. *Expert Systems with Applications*, 215:119307, 2023. [7](#)
- [50] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. [3](#), [8](#), [6](#)
- [51] Yabin Zhang, Bin Deng, Kui Jia, and Lei Zhang. Gradual domain adaptation via self-training of auxiliary models. *arXiv preprint arXiv:2106.09890*, 2021. [6](#)
- [52] Yukang Zhang and Hanzi Wang. Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2153–2162, June 2023. [2](#), [7](#), [5](#)
- [53] Yukang Zhang and Hanzi Wang. Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2153–2162, June 2023. [7](#)
- [54] Yukang Zhang, Yan Yan, Yang Lu, and Hanzi Wang. Towards a unified middle modality learning for visible-infrared person re-identification. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 788–796, 2021. [1](#), [3](#), [7](#)
- [55] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. [3](#)
- [56] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Deep domain-adversarial image generation for domain generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13025–13032, 2020. [3](#)
- [57] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. *arXiv preprint arXiv:2104.02008*, 2021. [3](#)
- [58] Kuan Zhu, Haiyun Guo, Zhiwei Liu, Ming Tang, and Jinqiao Wang. Identity-guided human semantic parsing for person re-identification. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 346–363. Springer, 2020. [2](#), [3](#)
- [59] Yuanxin Zhu, Zhao Yang, Li Wang, Sai Zhao, Xiao Hu, and Dapeng Tao. Hetero-center loss for cross-modality person re-identification. *Neurocomputing*, 386:97–109, 2020. [1](#), [3](#)

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

<sup>1</sup>Mahdi Alehdaghi, <sup>2</sup>Pourya Shamsolmoali, <sup>1</sup>Rafael M. O. Cruz, and <sup>1</sup>Eric Granger

<sup>1</sup>LIVIA, ILLS, Dept. of Systems Engineering, ETS Montreal, Canada

<sup>2</sup>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 $MI(P^1, \dots, P^K; Y)$

This section provides a proof that  $MI(P^1, \dots, P^K; Y)$  (Eq. 12) can be lower-bounded by :

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

following the properties of mutual information.

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

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

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

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

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

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

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

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

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

**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  $P^{k+1}, \dots, P^K$ . Using the **P.4** we have:

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

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

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

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

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

Given **P.1** and Eq. **5**,  $MI(P^1; P^2) = 0$  and  $MI(P^1; P^2; \tilde{P}^2) \leq MI(P^1; P^2) = 0$  so that:

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

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

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

To maximize  $\beta$ , we can rewrite and expand recursively in Eq. **A.7**:

$$MI(\tilde{P}^1; Y) = MI(P^2, \tilde{P}^2; Y) = \underbrace{MI(P^2; Y)}_{\text{maximizing}} + \underbrace{MI(\tilde{P}^2; Y)}_{\text{expanding}} - \underbrace{MI(P^2; \tilde{P}^2; Y)}_0. \quad (\text{A.12})$$

Therefore, it can be shown that:

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

and for  $k = 0$ , we have:

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

Finally, for maximizing  $MI(P^1, \dots, P^K; Y)$ , 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 $MI(P^k; Y)$

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: **13** of the manuscript). This approximation is formulated as **Proposition 1**.

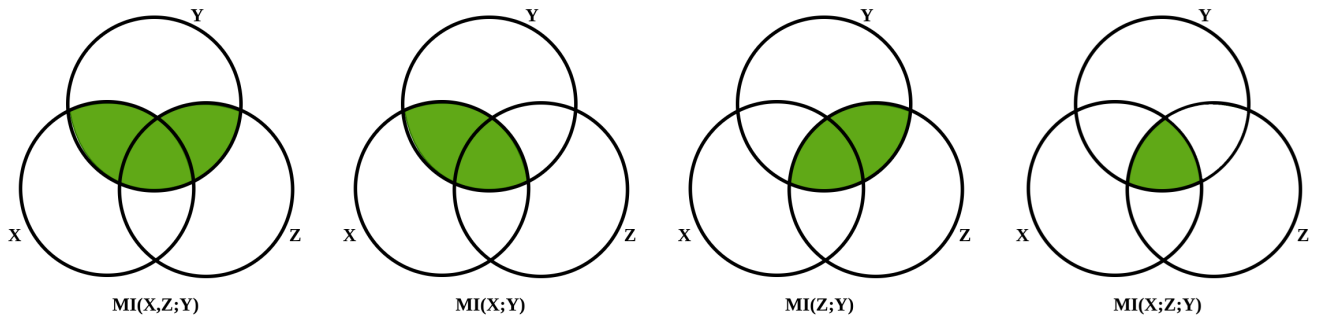


Figure A.1. Venn diagram of theoretic measures for three variables  $X$ ,  $Y$ , and  $Z$ , represented by the lower left, upper, and lower right circles, respectively.

**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,

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

Since the domain  $\mathcal{Y}$  does not change, the entropy of the identity  $\delta$  term is a constant and can therefore be ignored. Maximizing  $MI(P^k, Y)$  can only be achieved by minimizing the  $\xi$  term. We show that  $\mathcal{H}(Y|P^k)$  is upper-bounded by our cross-entropy loss (Eq. 13), and minimizing such loss results in minimizing the  $\xi$  term. By expanding its relation to the cross-entropy [?]:

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

where:

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

Through the minimization of Eq. 13, 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  $t = 1, \dots, T$  do ▷ over  $T$  steps
2:   while all batches are not selected do
3:      $x_v^j, x_i^j \leftarrow \text{batchSampler}(N_b, N_p)$ 
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. 21
16:   end while
17: end for

```

---

### B.2. Attentive Prototype Embedding

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

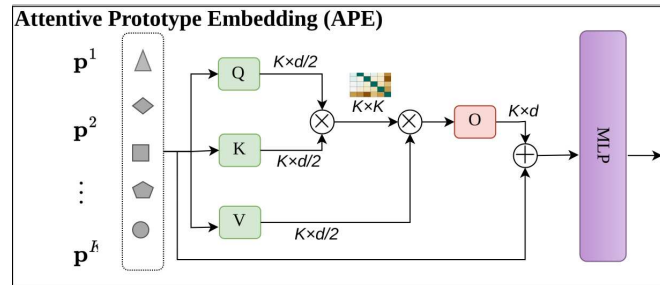


Figure B.1. Attentive prototype embedding (APE) architecture.

## Appendix C. Additional Details on the Experimental Methodology

### C.1. Datasets:

Research on cross-modal V-I ReID has extensively used the SYSU-MM01 [?], RegDB [?], 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 [?] 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  $\lambda_f$ ,  $\lambda_v$ ,  $\lambda_p$ , and  $\lambda_i$  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  $\lambda_f$ . As shown in Fig. D.1, accuracy improves with increasing  $\lambda_f$  until it reaches 0.1. Elevated  $\lambda_f$  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  $\lambda_v$  from 0.01 to 0.05, resulting in improved performance. Higher  $\lambda_v$  compresses prototype regions excessively, lacking sufficient ID-related information. Conversely,  $\lambda_i$  enhances the discriminative capabilities of prototypes in images. Balancing these factors, we find optimal values of 0.05 and 0.4 for  $\lambda_v$  and  $\lambda_i$ , respectively. Additionally, based on experimentation, we set  $\lambda_p = 0.2$  at the end of our analysis.

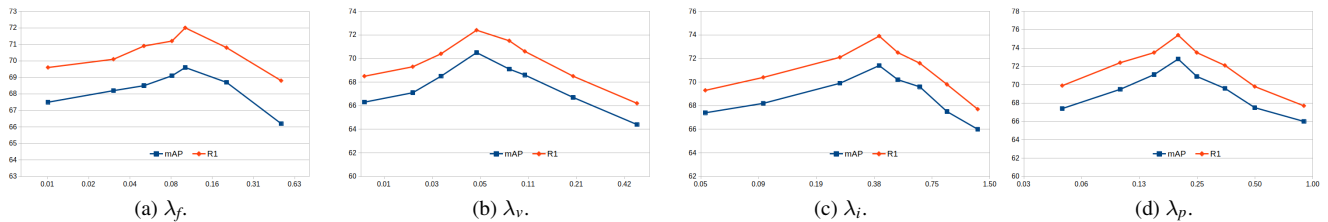


Figure D.1. Accuracy of the proposed BMDG over  $\lambda_f$ ,  $\lambda_v$ ,  $\lambda_i$ , and  $\lambda_p$  values on SYSU-MM01 dataset in all-search and single-shot mode.

### 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 [?] setting for different numbers of part prototypes ( $K$ ) and intermediate steps ( $T$ ).

(a) Prototype exchanging							(b) Mixup [?]						
$T$	Number of part prototypes ( $K$ )						$T$	Number of part prototypes ( $K$ )					
	3	4	5	6	7	10		3	4	5	6	7	10
0	65.98	67.03	67.23	68.11	67.55	65.66	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	1	66.31	67.22	67.31	68.28	68.5	65.8
2	69.51	70.08	70.72	71.14	69.97	71.25	4	66.50	67.68	67.79	68.52	68.49	65.88
3	71.44	71.69	71.82	72.02	71.06	71.67	6	66.98	67.77	68.05	<b>68.73</b>	68.56	66.02
4	-	71.98	72.15	<b>72.86</b>	71.19	69.00	10	67.04	67.60	67.93	68.65	68.54	65.57
6	-	-	-	72.40	71.17	69.54							
10	-	-	-	-	-	69.46							

### 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 [?]	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 [?] 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



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 [?]. Each color shows the identity.

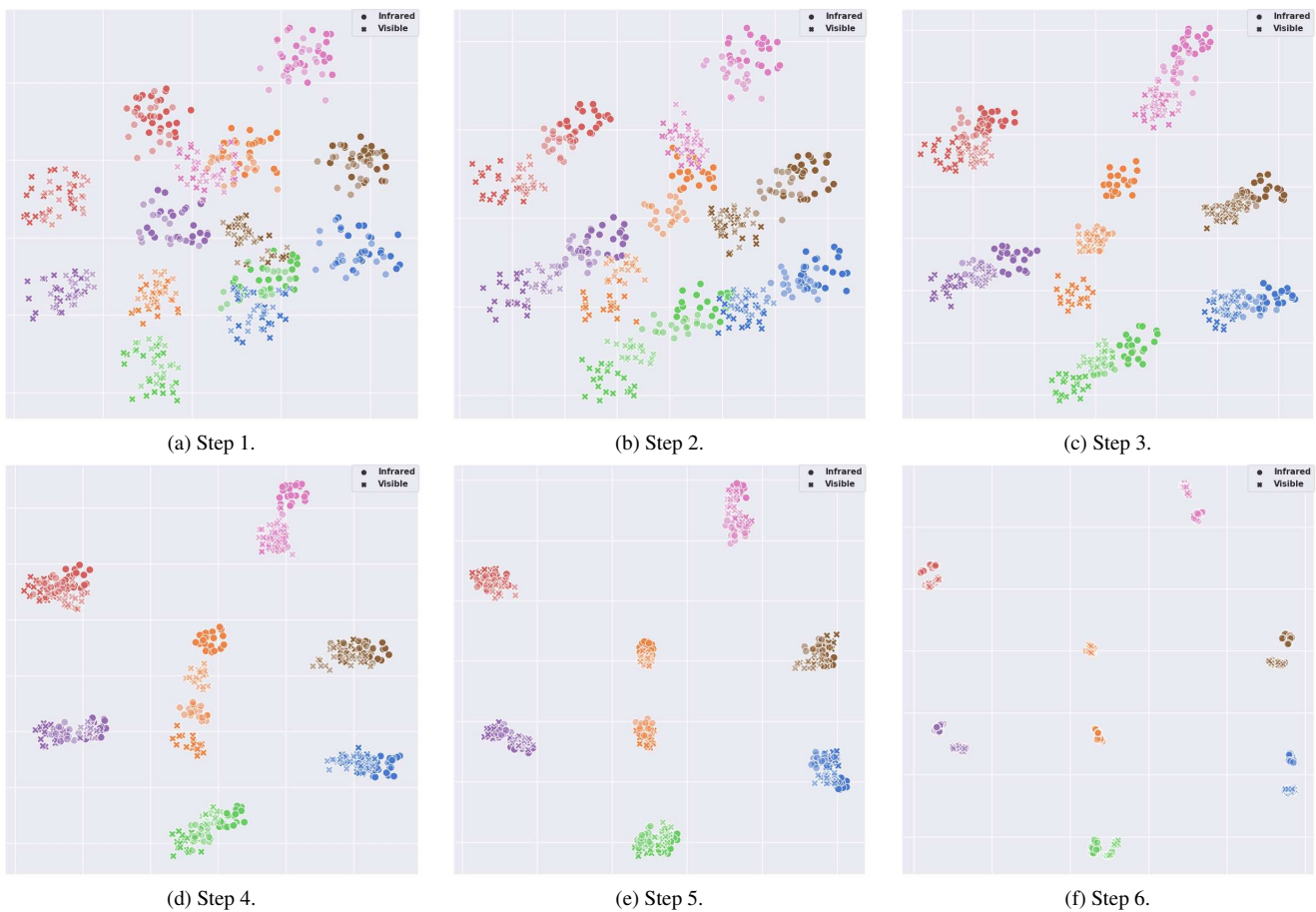
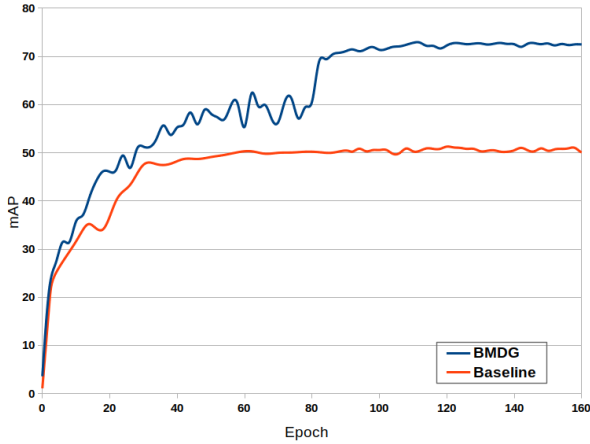
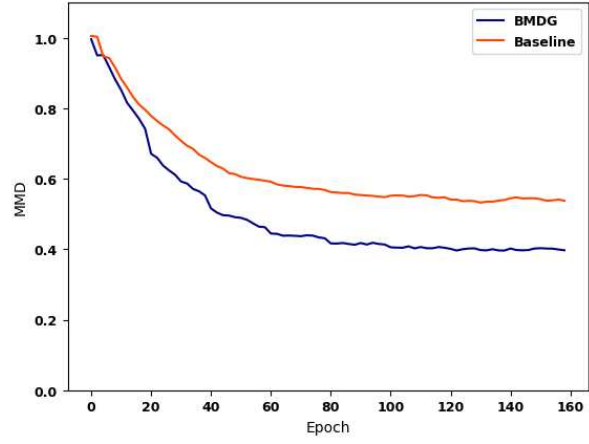


Figure E.2. Distributions of learned  $V$  and infrared features of 7 identities from SYSU-MM01 dataset in training for 6 steps at epochs of 10,30,50,70,90 and 160 respectively by UMAP [?]. Each color shows the identity. The intermediate features are drawn with lower opacity.

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 [?]. (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  $18 \times 9$ , 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 [?] on the PASCAL-Part Dataset [?] as the teacher. In Fig. E.5, the results show our model’s strong capacity for detecting body parts compared to its teacher.

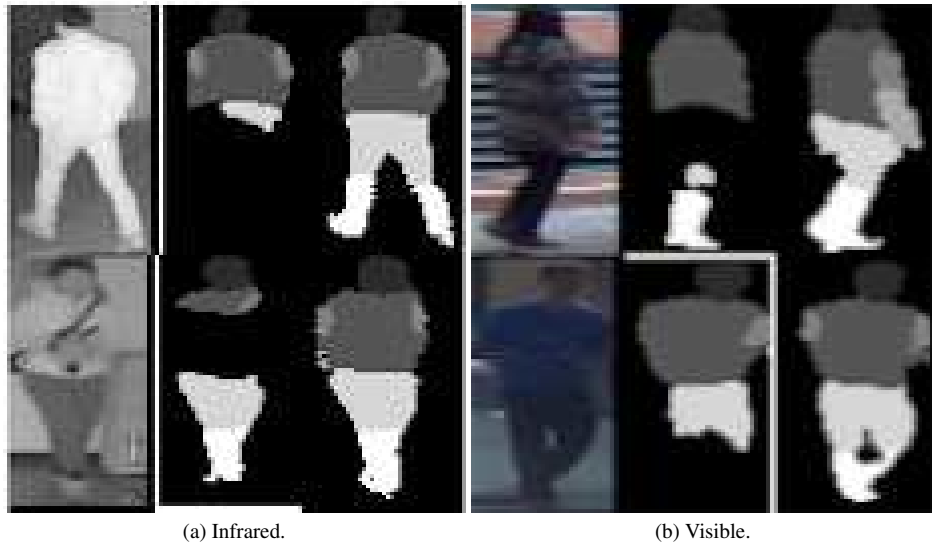


Figure E.5. Semi-supervised part discovery on the SYSU-MM01 dataset. The first columns are the (a) infrared and (b) visible images. The second column images are the result from [?], and the last column are results with our fine-tuned model in BMDG.