Hybrid Modality Metric Learning for Visible-Infrared Person Re-Identification

LA ZHANG, Beijing Institute of Technology, China
HAIYUN GUO and KUAN ZHU, Institute of Automation Chinese Academy of Sciences, China
HONGLIN QIAO and GAOPAN HUANG, Alibaba Cloud, China
SEN ZHANG and HUICHEN ZHANG, Traffic Management Research Institute of the Ministry of Public Security, China
JIAN SUN, Beijing Institute of Technology, China
JINQIAO WANG, Institute of Automation Chinese Academy of Sciences, China

Visible-infrared person re-identification (Re-ID) has received increasing research attention for its great practical value in night-time surveillance scenarios. Due to the large variations in person pose, viewpoint, and occlusion in the same modality, as well as the domain gap brought by heterogeneous modality, this hybrid modality person matching task is quite challenging. Different from the metric learning methods for visible person re-ID, which only pose similarity constraints on class level, an efficient metric learning approach for visible-infrared person Re-ID should take both the class-level and modality-level similarity constraints into full consideration to learn sufficiently discriminative and robust features. In this article, the hybrid modality is divided into two types, within modality and cross modality. We first fully explore the variations that hinder the ranking results of visible-infrared person re-ID and roughly summarize them into three types: within-modality variation, cross-modality modality-related variation, and cross-modality modality-unrelated variation. Then, we propose a comprehensive metric learning framework based on four kinds of paired-based similarity constraints to address all the variations within and cross modality. This framework focuses on both class-level and modality-level similarity relationships between person images. Furthermore, we demonstrate the compatibility of our framework with any paired-based loss functions by giving detailed implementation of combing it with triplet loss and contrastive loss separately. Finally, extensive experiments of our approach on SYSU-MM01 and RegDB demonstrate the effectiveness and superiority of our proposed metric learning framework for visible-infrared person Re-ID.

This work was supported by Key-Area Research and Development Program of Guangdong Province (No. 2020B010165001), National Natural Science Foundation of China (No. 61772527, 62002356, 61925303), Open Project of Key Laboratory of Ministry of Public Security for Road Traffic Safety (No. 2020ZDSYSKFKT04).

Authors' addresses: L. Zhang, Beijing Institute of Technology, No. 5, South Street, Zhongguancun, Haidian District, Beijing, Haidian Qu, Beijing Shi, China, 100081; email: 3120185467@bit.edu.cn; H. Guo (corresponding author), K. Zhu, and J. Wang, Institute of Automation Chinese Academy of Sciences, 95 Zhongguancun East Road, Haidian Qu, Beijing Shi, China; emails: {haiyun.guo, kuan.zhu, jqwang}@nlpr.ia.ac.cn; H. Qiao, Alibaba Cloud, Radiance JinHui Tower, Building No. 6, The 4th District, Wangjing DongYuan, Chaoyang District, Beijing, China; email: kenny.qhl@alibaba-inc.com; G. Huang, Alibaba Cloud, Ali Center, Nanjing, Jiangsu, China; email: huanggpan@163.com; S. Zhang and H. Zhang, Traffic Management Research Institute of the Ministry of Public Security, 88 Qianrong Road, Wuxi, Jiangsu Sheng, China; emails: {76806915, 496306636}@qq.com; J. Sun, Beijing Institute of Technology, No. 5, South Street, Zhongguancun, Haidian District, Beijing, Haidian Qu, Beijing Shi, China; email: sunjian@bit.edu.cn.

1551-6857/2022/01-ART25 $15.00
https://doi.org/10.1145/3473341

CCS Concepts: • **Computing methodologies** → **Artificial intelligence**; • **Artificial intelligence** → *Computer vision*; • **Computer vision** → Computer vision tasks; • **Computer vision tasks** → Biometrics;

Additional Key Words and Phrases: Visible-infrared person re-identification, cross-modality, metric learning

ACM Reference format:
La Zhang, Haiyun Guo, Kuan Zhu, Honglin Qiao, Gaopan Huang, Sen Zhang, Huichen Zhang, Jian Sun, and JinQiao Wang. 2022. Hybrid Modality Metric Learning for Visible-Infrared Person Re-Identification. *ACM Trans. Multimedia Comput. Commun. Appl.* 18, 1s, Article 25 (January 2022), 15 pages.
https://doi.org/10.1145/3473341

1 INTRODUCTION

Given the query person represented by an infrared image (or a RGB image), **visible-infrared person re-identification (Re-ID)** aims to retrieve other RGB images (or infrared images) involving the identical person from the gallery dataset [22, 38]. As we know, night is high incidence of security incidents, while traditional video surveillance device or technology does not work well at night due to the limitations of visible cameras in poor illumination conditions, e.g., in the night or dark. To compensate for the shortcomings of visible cameras, infrared cameras are widely adopted in video surveillance systems. But the infrared images fail to capture many critical visual clues for person identification, such as clothing color. The combination of visible and infrared person images has been an essential solution to improve the person Re-ID performance in weak illumination. Currently, more and more cameras installed in cities can support both visible and infrared imaging, which provides convenient infrastructure conditions for wide application of this technology.

In a visible person Re-ID task, image discrepancies of the same pedestrian (intra-class variation) are already quite large because of variable human poses, camera viewpoints, and variations of illumination, which can be even larger than those of different pedestrians. Moreover, when modalities with quite different imaging principles are introduced, the variations of intra-class and inter-class could become more complex. Hence, it is quite difficult to learn efficient feature representations with semantic consistency for the identical person. The learned feature representation should effectively capture the identifiable person characteristics contained both in visible and infrared images as well as possess sufficient robustness against intra-class variations both within and cross modality.

Many deep metric learning methods have been proposed for visible person Re-ID to learn efficient feature embedding space where both intra-class compactness and inter-class discrepancy are enhanced [3, 12, 14, 20, 25, 26, 28, 29, 31]. Regardless of the loss function they adopt, these methods only need to pose class-level similarity constraints on the feature learning network. However, due to the additional challenge caused by heterogeneous modality, an efficient metric learning method for visible-infrared person Re-ID should consider both the class-level and modality-level similarity constraints. In this article, we first divide the hybrid modality into two types, within modality and cross modality. And then we fully explore the variations existing in this cross-modality person matching task and roughly summarize them into three variations as illustrated in Figure 1. Given a triplet of image samples, i.e., a query image, a positive that shares the same person identity with the anchor and a negative that belongs to a different personal identity, we define the following variations:

- Within-modality Variation. Within modality is when all samples come from the same modality. Variations within modality are mainly caused by variable human poses, camera viewpoints, and variations of illumination, which may result in small intra-class similarity and

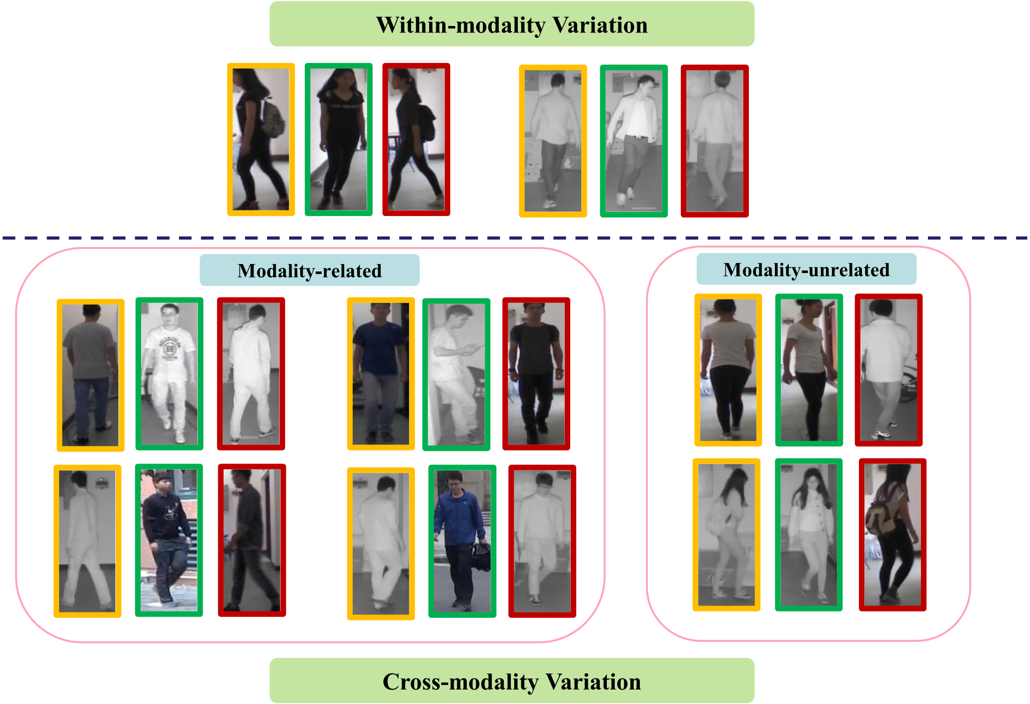


Fig. 1. Illustration of all types of variations that hinder the visible-infrared person Re-ID results. For each triplet of images, the image in the yellow box denotes the query, the one in green shares the same identity with the query, and the one in red belongs to another identity. In the ideal feature space, the feature for the image in green should be closer than that for red to the query. However, due to the interference of illustrated variations, the similarity relationship is broken in the returned Re-ID results. Within-modality variation corresponds to variables such as person pose, occlusion, and so on. Cross-modality variation can be further divided into modality-related and modality-unrelated variations.

large inter-class similarity. The single-modality person Re-ID only needs to address this type of variation.

- Cross-modality Modality-related Variation. For a triplet of samples coming from different modalities, we define the variations that may disturb their similarity relationship in the result as the cross-modality variation. And we further divide it into the modality-related and the modality-unrelated variations based on whether the variation is brought by heterogeneous modality or not. As illustrated in the left bottom of Figure 1, the huge modality gap makes it difficult for the feature embedding to pull the distance between positive RGB-Infrared image pairs closer, let alone make it even closer than that of negative RGB-RGB (or Infrared-Infrared) image pairs.
- Cross-modality Modality-unrelated Variation. Apart from the modality gap, the modality-unrelated variation can also disrupt the similarity ranking result for the cross-modality person matching. As illustrated in the right bottom of Figure 1, the large pose variation makes the distance between positive RGB-RGB image pair even larger than that of negative RGB-Infrared image pair. This variation corresponds to the extreme condition of within-modality variation, where the variation caused by modality-unrelated factors is even larger than cross-modality discrepancy. We define this variation apart from the within-modality variation to emphasize this extreme condition.

Earlier studies concentrate on class-level feature learning with metric learning methods of visible person Re-ID [33]. They limit the discriminability of feature representations without considering modality-specific information. In later research, researchers begin to consider the impact of modality-level similarity constraints. Ye et al. [34, 35] adopt triplet loss for within-modality variation and part of cross-modality modality-related variations, named BDTR, which is proved to be an efficient way to learn more discriminative feature representation. In most subsequent studies, combing **Cross-Entropy loss (CE loss)** and ranking loss becomes a general loss function choice to tackle within-modality variations or part of cross-modality variations or both of them [5, 13, 34, 37]. However, most of them do not fully explore the cross-modality variations. **Hypersphere Manifold Embedding (HSME)** [13] takes cross-modality modality-related variations into consideration but still underlooks cross-modality modality-unrelated variations. Hence, the existing works ignore some of the aforementioned three variations more or less, which are considered to be suboptimal.

To address the above issue, this article proposes a **hybrid modality metric learning framework (HMML)** for visible-infrared person Re-ID, which involves four kinds of paired-based similarity constraints to tackle all the above within and cross modality variations. Specifically, we first construct a within-modality similarity constraint to enforce the intra-class similarity and inter-class discriminativeness of the learned features against the within-modality variations. Then we propose the cross-modality modality-unrelated similarity constraint to further reduce the large intra-class variation within modality. Additionally, we propose two kinds of cross-modality modality-related similarity constraints to further close the gap between heterogeneous modalities. Combing the above four similarity constraints together, we can effectively guide the network to learn more robust and discriminative feature representations for visible-infrared person Re-ID. Furthermore, to demonstrate the generalization of our metric learning framework, we introduce it into the construction of two classical supervision loss functions, i.e., triplet loss [14] and contrastive loss [8, 29]. Besides, our framework is also compatible with other paired-based loss functions such as n-pair loss [27]. Finally, we conduct extensive experiments to demonstrate the effectiveness of cross-modality modality-unrelated term and one kind of cross-modality modality-related terms and investigate the influence of hyper-parameters for these two modules. Compared with the state-of-the-art metric learning methods adopted by existing visible-infrared approaches, our method promotes the best performance by 6.7% on SYSU-MM01 and 7.73% on RegDB in terms of Rank-1.

The main contributions can be summarized as follows:

- We fully explore the complex within and cross modality variations for visible-infrared person Re-ID and propose a comprehensive hybrid modality metric learning framework based on both class-level and modality-level similarity constraints to tackle all types of variations
- We demonstrate the compatibility of our framework with any paired-based loss functions by giving detailed implementation of combing it with triplet loss and contrastive loss separately.
- Extensive experiments of our approach on SYSU-MM01 and RegDB demonstrate the effectiveness and superiority of our approach compared with other metric learning methods for visible-infrared person Re-ID.

2 RELATED WORK

Visible-infrared person Re-ID is first defined by Wu et al. [33]. They propose a deep zero-padding method to learn shared feature space for visible-infrared matching by training a one-stream network; meanwhile, they release a large-scale visible-infrared person Re-ID dataset, named SYSU-MM01. However, the learned feature representations discriminability is limited without considering modality-specific information. Then Ye et al. [34] propose a two-stage framework (termed TONE). TONE contains both feature learning and metric learning. The parameters of the shallow layers for two modalities are independent to extract the model-specific feature representations, and parameters of deep layers are shared to extract model-shared feature representations [34, 35]. Yi et al. [13] propose a dual-stream hypersphere manifold embedding network (named D-HSME) by using both CE loss and ranking loss. The loss function constrain some of cross-modality modality-related variations and within-modality variations. A dual-path feature learning framework named DSCSN [36] first embeds input image pairs into a three-dimensional tensor space and then extracts contrastive features by comparing positive and negative pairs dynamically. A Pairwise Binary Cross-Entropy loss and CE loss are used to supervise the learning process. Ye et al. [23] propose a **modality-aware collaborative learning method (MAC)**, which handles the modality-discrepancy in both feature level and classifier level. **The Two-Stream Local Feature Network (TSLFN)** [1] concentrates on reducing intra-class variations by pulling in the distance of the intra-class cross-modality center, called Hetero-Center loss. Jia et al. [16] proposes a **similarity inference metric (SIM)**, which is used to circumvent the cross-modality discrepancy by exploits intra-modality sample similarities.

Due to the excellent performance of **generative adversarial networks (GAN)** [8, 10, 15, 24, 32, 40] in other research areas, some researchers try to apply GAN to visible-infrared person Re-ID. The method named cmGAN [5] is a kind of cross-modality generative adversarial network to handle the lack of insufficient discriminative information, and the metric learning integrates both identification loss and cross-modality triplet loss. The **Dual-level Discrepancy Reduction Learning ($\mathbf{D^2}$RL)** [39] uses domain-specific image generation to tackle the modality discrepancy. Wang et al. [30] point at that most of current methods focus on set-level information, infrared sets and BGR sets, which lacks attention for instance information. They propose a method named **Joint Set-level and Instance-Level Alignment Re-ID (JSIA)** to generate visible-infrared paired images, which are used to handle both set-level and instance-level alignments. And always there are other efficient ways to handle visible-infrared person Re-ID. **The X-Infrared-Visible Re-ID cross-modal learning framework (XIV)** [17] conducts a third modality transferring both visible and infrared modalities to a unified feature space. Seokeon et al. [4] disentangle ID-discriminative factors and ID-excluded factors by a generation network with a hierarchical feature learning module (named Hi-CMD), and the ID-discriminative feature is used for robust visible-infrared pedestrian matching. Lu et al. [21] aim to learn discriminative and complementary shared and specific features by proposing a cross-modality shared-specific feature transfer algorithm (named cm-SSFT), which first extracts shared features by a two-stream network with an extra module of modality adaption, project adversarial learning and reconstruction blocks.

Above all, in the field of infrared-visible person Re-ID, it is quite difficult to learn discriminative feature representations due to the complex variations. To solve this problem, some researchers focus on metric learning [1, 13, 16, 23, 34–36], and some researchers try other methods [4, 5, 17, 21, 30, 39]. Our approach focus on within- and cross-modality metric learning to address all complex variations based on exploring similarity relationships for image pairs.

3 PROPOSED METHOD

We construct four kinds of similarity constraints to overcome the aforementioned hybrid modality variations on both class level and modality level as shown in Figure 1. We gain a comprehensive metric learning framework combining these four constraints, which could result in a better Re-ID result. Based on our framework, we can upgrade any pair-based metric learning methods to the version considering all variations, and then these optimized metric learning methods can be used to supervise CNN network to learn more robust and discriminative feature representations. In this section, we first describe our HMML and then give a detailed implementation of combing HMML with triplet loss and contrastive loss separately.

3.1 Proposed Metric Learning Framework with Both Class-level and Modality-level Similarity Constraints

Most of the existing metric learning methods for visible-infrared person Re-ID only consider the class-level similarity constraint for the overall image data, which overlooks the impact of modalities differences [7, 33]. Later studies demonstrate that modality-level similarity constraints achieve better performance in visible-infrared person Re-ID [16, 34, 37]. However, existing research does not make a thorough exploration of the complex variations that result in disturbed similarity ranking relationships for this cross-modality person retrieval task. This article shows a detailed analysis of these complex variations, and we roughly divide them into three types: within-modality variations, cross-modality modality-related variations, and cross-modality modality-unrelated variations, in terms of within or cross modality and modality related or unrelated. For targeting these issues, we propose different similarity constraints to tackle them. The metric learning based on each of them is a kind of local metric learning, and then we integrate all of them into a comprehensive metric learning framework. This framework concentrates on constraining positive pairs smaller than negative pairs, regardless of hybrid modality variations.

For describing our approach, we use V and T to represent the RGB image set and the infrared image set. Let $x \in V$ denotes RGB images and $z \in T$ denotes infrared images. Given a visible image x (or an infrared image z) of a specific person, x_p (z_p) and x_n (z_n) represent positive images and negative images. We denote the feature extractor as $f(\cdot)$, the extracted features of x and z are represented by $f(x)$ and $f(z)$. The Euclidean distance $D(\cdot,\cdot)$ is used for measuring the similarity of image pairs, e.g., the Euclidean distance $D(x,z)$ between x and z is represented by

$$D(x,z) = \|f(x) - f(z)\|_2 . \tag{1}$$

Given a set of image $\{x, x_n, x_p, z_n, z_p\}$, the following four types of constraints are defined to address all the variations discussed in Section 1:

- Within-modality Similarity Constraint (WM constraint). This constraint aims at tackling the within-modality variations and enforce that feature distance between positive pairs should be smaller than that for negative pairs:

$$D(x,x_p) < D(x,x_n), D(z,z_p) < D(z,z_n). \tag{2}$$

- Cross-modality Modality-unrelated Similarity Constraint (CM_U constraint). By forcing the feature distance between positive image pairs from the same modality to be smaller than that of negative image pairs from different modalities, this constraint focus on the relatively large within-modality variations during the metric learning process, which can be expressed as follows:

$$D(x,x_p) < D(x,z_n), D(z,z_p) < D(z,x_n). \tag{3}$$

- Cross-modality Modality-related Similarity Constraint. As discussed in Section 1, visible-infrared Re-ID also suffers from large cross-modality variations. To address this issue, we further divide this constraint into two sub-modules. One module focus on pushing the distance of negative pairs within a modality (RGB-RGB or Infrared-Infrared) larger than the distance of positive image pairs from different modalities (RGB-Infrared). In the other module, given a specific sample, the positive and negative are all from another modality, which is different from itself. The former constraint is mainly caused by the similarity within a modality (CM_S constraint). The latter constraint pays more attention to the gap of different modalities (CM_G constraint), which focuses on reducing modality-level discrepancy by guaranteeing the consistency of similarity ranking relationships in cross-modality. CM_S

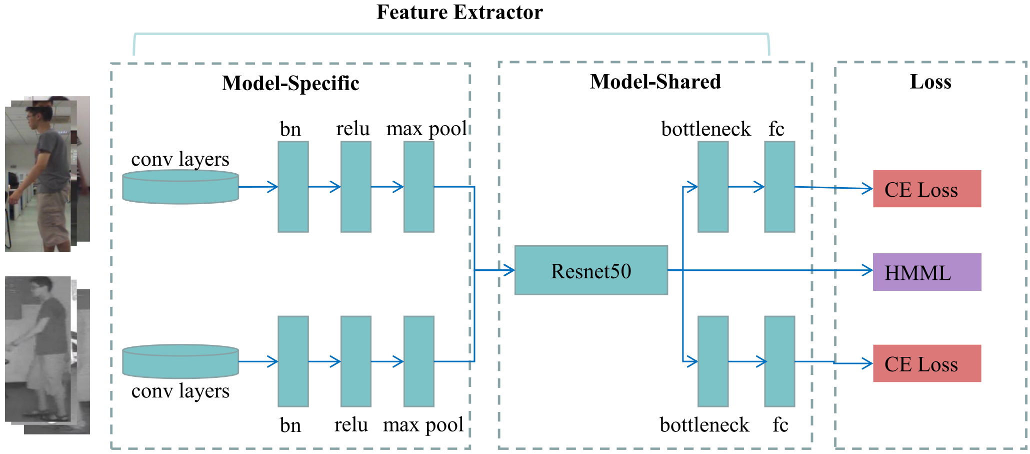


Fig. 2. The overall network architecture of our proposed visible-infrared person Re-ID method based on within- and cross-modality metric learning. A series of cross-modality images are input into the network. The feature extractor learns model-specific features first and then concatenates two features to learn model-shared characteristics. Model-shared features are trained by our framework with triplet loss or contrastive loss. Model-specific features are trained by CE loss.

constraint and CM_G constraint are expressed by Equations (4) and (5) as follows:

$$D(x, z_p) < D(x, x_n), D(z, x_p) < D(z, z_n), \tag{4}$$

$$D(x, z_p) < D(x, z_n), D(z, x_p) < D(z, x_n). \tag{5}$$

Above all, our proposed HMML consists of four constraints, WM constraint, CM_U constraint, CM_S constraint, and CM_G constraint. We adopt a dual-path end-to-end learning framework [22, 35]. The overall network architecture of our proposed method is as shown in Figure 2. First, a feature extractor is used to learn model-specific features, and then we concatenate two model-specific features to learn model-shared features. HMML is used to constrain model-shared features. To overcome the slow convergence of pair-based loss, CE loss is always integrated. With the supervision of HMML, the overall network can learn more discriminative feature representations tackling all types of variations.

3.2 Compatibility Analysis with Pair-based Loss Functions

Generally, we embed a group of images to a feature space via a feature extractor first and then define a metric function to measure the similarity between image pairs. Based on comprehensive analysis for pair-based similarity in visible-infrared person Re-ID, our proposed HMML framework is compatible with regular metric learning, i.e., contrastive loss, triplet loss, N-Pair loss, and so on. Among them, triplet loss and contrastive loss are most widely used in person Re-ID. In this section, we give a detailed implementation of combing HMML with contrastive loss and triplet loss separately.

In this section, given a mini-batch, it contains N visible and N infrared images. $\{x_i\}_{i=1}^{N} \in V$ are visible samples and $\{z_i\}_{i=1}^{N} \in T$ are infrared samples. y_i represents the label of x_i or z_i, and we hypothesize the subscript of y, i, and j mean same pedestrian and k is different from i and j $(y_i = y_j, y_i \neq y_k)$.

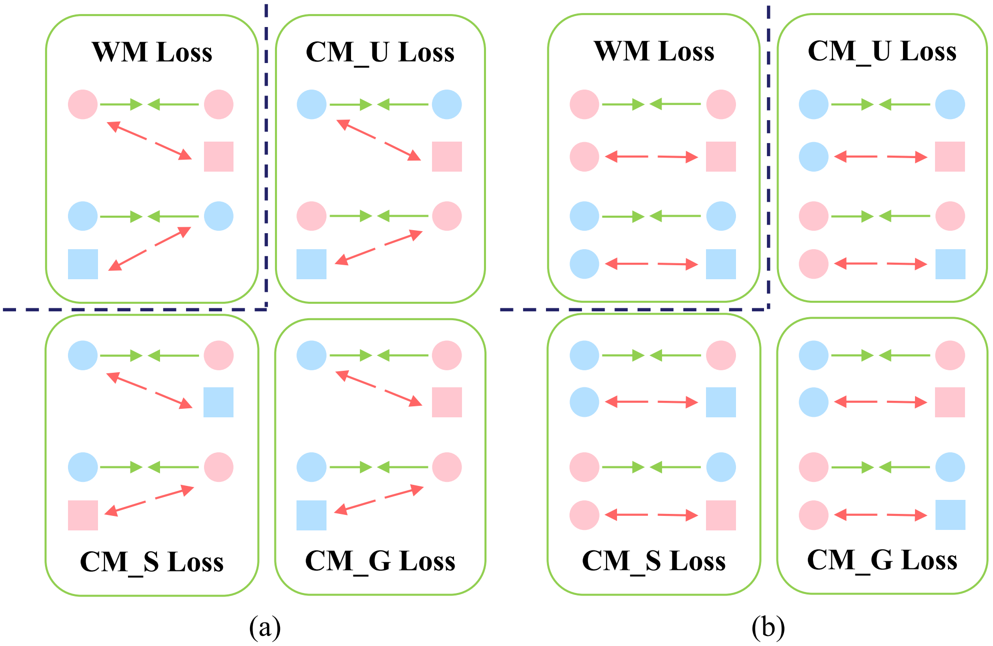


Fig. 3. (a) Our proposed HMML framework with triplet loss. Each of the loss terms is used to forcing the distance between anchor to its positive is smaller than the anchor to its negative. (b) Our proposed HMML framework with contrastive loss. Each of the loss terms is used to constrain positive pairs to be closer and push away negative pairs. In this figure, different colors represent different modalities and different shapes represent different persons. Green arrows represent pull in the notes and red arrows represent push away the notes.

3.2.1 HMML with Contrastive Loss. The main idea of contrastive loss is that the distance of positive pairs tends to zero; meanwhile, the distance of negative pairs should be larger than a predefined margin [8, 22, 29]. As illustrated in Figure 3(b), WM Loss is used for closing within-modality positive image pairs $\{x_i, x_j\}$ ($\{z_i, z_j\}$), and push away within-modality negative image pairs $\{x_i, x_k\}$ ($\{z_i, z_k\}$). We use ρ_1 to represent the predefined margin, and the loss of WM constraint is formulated by:

$$L_{WM\_C} = \sum_{\forall y_i = y_j} D(x_i, x_j) + \sum_{\forall y_i \neq y_k} [\rho_1 - D(x_i, x_k)]_+ + \sum_{\forall y_i = y_j} D(z_i, z_j) + \sum_{\forall y_i \neq y_k} [\rho_1 - D(z_i, z_k)]_+. \quad (6)$$

As discussed in the Section 3.1, *CM_U* constraint addresses the extreme condition of within-modality variation. This loss function supervises the within-modality positive pairs $\{x_i, x_j\}$ ($\{z_i, z_j\}$) become more similar, and cross-modality negatives $\{x_i, z_k\}$ ($\{z_i, x_k\}$) become more discriminative. *CM_U* constraint with contrastive loss is formulated by

$$L_{CM\_U\_C} = \sum_{\forall y_i = y_j} D(x_i, x_j) + \sum_{\forall y_i \neq y_k} [\rho_1 - D(x_i, z_k)]_+ + \sum_{\forall y_i = y_j} D(z_i, z_j) + \sum_{\forall y_i \neq y_k} [\rho_1 - D(z_i, x_k)]_+. \quad (7)$$

The *CM_S* constraint concentrates on overcoming the similarity within a modality. The loss function guarantees the cross-modality positive pair $\{x_i, z_j\}$ ($\{z_i, x_j\}$) to be closer and pushes away the within-modality negative pair $\{x_i, x_k\}$ ($\{z_i, z_k\}$). The *CM_S* constraint with contrastive loss is formulated by

$$L_{CM\_S\_C} = \sum_{\forall y_i = y_j} D(x_i, z_j) + \sum_{\forall y_i \neq y_k} [\rho_1 - D(x_i, x_k)]_+ + \sum_{\forall y_i = y_j} D(z_i, x_j) + \sum_{\forall y_i \neq y_k} [\rho_1 - D(z_i, z_k)]_+. \tag{8}$$

The CM_G constraint with contrastive loss is used to keep the cross-modality positive pair $\{x_i, z_j\}$ ($\{z_i, x_j\}$) closer, while pushing away the cross-modality negative pair $\{x_i, z_k\}$ ($\{z_i, x_k\}$), as follows:

$$L_{CM\_G\_C} = \sum_{\forall y_i = y_j} D(x_i, z_j) + \sum_{\forall y_i \neq y_k} [\rho_1 - D(x_i, z_k)]_+ + \sum_{\forall y_i = y_j} D(z_i, x_j) + \sum_{\forall y_i \neq y_k} [\rho_1 - D(z_i, x_k)]_+. \tag{9}$$

In addition, λ_{WM}, λ_{CM_U}, λ_{CM_S}, and λ_{CM_G} are tradeoff parameters for the corresponding loss terms separately. We use L_{ID} to represent the CE loss, and α_{ID} is the tradeoff parameter for it. The overall implementation of our framework based on contrastive loss is as follows:

$$L_C = \lambda_{WM} L_{WM\_C} + \lambda_{CM\_U} L_{CM\_U\_C} + \lambda_{CM\_S} L_{CM\_S\_C} + \lambda_{CM\_G} L_{CM\_G\_C} + \alpha_{ID} L_{ID}. \tag{10}$$

3.2.2 HMML with Triplet Loss. Triplet loss is designed based on a series of triplets, which contains one anchor sample, one positive sample with the same identity, and one negative sample from a different identity [2]. The main idea is that the distance between anchor to its positive is smaller than the anchor to its negative by a pre-defined margin [13, 14]. Based on triplet loss, we adopt the hard sample mining strategy in default.

As illustrated in Figure 3(a), WM constraint focus on forcing the max distance between x_i and x_j (z_i and z_j) smaller than the min distance of x_i and x_k (z_i and z_k) by a predefined margin ρ_2. WM constraint with triplet loss is as follows:

$$L_{WM\_T} = \sum_{\forall y_i = y_j} max[\rho_2 + D(x_i, x_j) - \min_{\forall y_i \neq y_k} D(x_i, x_k), 0] + \sum_{\forall y_i = y_j} max[\rho_2 + D(z_i, z_j) - \min_{\forall y_i \neq y_k} D(z_i, z_k), 0]. \tag{11}$$

CM_U constraint is used to ensure that the distance of anchor to its furthest within-modality positive samples x_i and x_j (z_i and z_j) is smaller than the anchor to its nearest cross-modality negative samples x_i and z_k (z_i and x_k). CM_U constraint with triplet loss is as follows:

$$L_{CM\_U\_T} = \sum_{\forall y_i = y_j} max[\rho_2 + D(x_i, x_j) - \min_{\forall y_i \neq y_k} D(x_i, z_k), 0] + \sum_{\forall y_i = y_j} max[\rho_2 + D(z_i, z_j) - \min_{\forall y_i \neq y_k} D(z_i, x_k), 0]. \tag{12}$$

CM_S variation may result in that the distance of anchor to its furthest cross-modality positive samples x_i and z_j (z_i and x_j) is larger than the anchor to its nearest within-modality negative samples x_i and x_k (z_i and z_k). Hence, the CM_S constraint with triplet loss is used to address this issue, and it is formulated by

$$L_{CM\_S\_T} = \sum_{\forall y_i = y_j} max[\rho_2 + D(x_i, z_j) - \min_{\forall y_i \neq y_k} D(x_i, x_k), 0] + \sum_{\forall y_i = y_j} max[\rho_2 + D(z_i, x_j) - \min_{\forall y_i \neq y_k} D(z_i, z_k), 0]. \tag{13}$$

CM_G constraint focuses on narrowing the gap of different modalities by handling the distance of anchor to its furthest cross-modality positive sample x_i and z_j (z_i and x_j) is smaller than the anchor to its nearest cross-modality negative sample x_i and z_k (z_i and x_k), the loss is formulated by

$$L_{CM\_G\_T} = \sum_{\forall y_i = y_j} max[\rho_2 + D(x_i, z_j) - \min_{\forall y_i \neq y_k} D(x_i, z_k), 0] + \sum_{\forall y_i = y_j} max[\rho_2 + D(z_i, x_j) - \min_{\forall y_i \neq y_k} D(z_i, x_k), 0]. \tag{14}$$

Above all, the overall triplet loss implementation with our framework is as follows:

$$L_T = \lambda_{WM} L_{WM\_T} + \lambda_{CM\_U} L_{CM\_U\_T} + \lambda_{CM\_S} L_{CM\_S\_T} + \lambda_{CM\_G} L_{CM\_G\_T} + \alpha_{ID} L_{ID}. \quad (15)$$

4 EXPERIMENTS

In this section, first, we introduce our implementation details, and then ablation experiments are conducted to validate the effectiveness of CM_S and CM_U terms in our HMML framework and to analyze the influence of hyper-parameters for these two terms. The ablation experiments are based on the implementation of HMML with triplet loss (HMML_T). We also analyze the performance of both implementations of HMML with triplet loss and contrastive loss (HMML_C), which shows the compatibility of HMML for typical metric learning. Finally, we compare the performance of HMML_T and HMML_C with state-of-the-art methods, which demonstrates the competitive performance of our approach on the SYSU-MM01 and RegDB datasets.

4.1 Implementation Details

Datasets. Our experiments are evaluated based on two public datasets, the RegDB dataset [7] and the SYSU-MM01 dataset [33].

RegDB involves 412 persons, there are 10 infrared images and RGB images for each person. Following the evaluation protocol in the research of Ye et al. [35], we randomly split the dataset into two halves, where one is for training and the other one is for testing. The final result is calculated by averaging the testing values of repeating the procedure 10 times.

SYSU-MM01 involves 287,628 RGB images and 15,792 infrared images from 4 RGB cameras and 2 infrared cameras. It contains 491 persons, among them, 296 persons for training, 99 persons for validation, and 96 persons for testing. We adopt the single-shot all-search mode evaluation protocol on this dataset [7].

Evaluation Metrics. Cumulated Matching Characteristics (CMC) curve, **Mean Average Precision (mAP)**, and **Mean Inverse Negative Penalty (mINP)** are adopted on both of SYSU-MM01 dataset and RegDB dataset to evaluate the effectiveness of methods.

Implementation Details. Our experiments implement with Pytorch. We adopt a two-stream network architecture, and the backbone uses renet50. It contains 64 images in each mini-batch, which contains 32 RGB images and 32 infrared images, and each modality contains 8 persons with 4 different human poses. Random cropping, random horizontal flip, and random erasing are utilized for data argumentation. We set the tradeoff parameters as $\lambda_{WM} = 0.1$, $\lambda_{CM\_U} = 0.1$, $\lambda_{CM\_G} = 1$, and $\lambda_{CM\_S} = 0.5$. The settings of λ_{WM} and λ_{CM_G} follow the research of Ye et al. [34], and settings for λ_{CM_S} and λ_{CM_U} will be mainly discussed in Section 4.2. The predefined margin ρ_1 and ρ_2 are all set to 0.3.

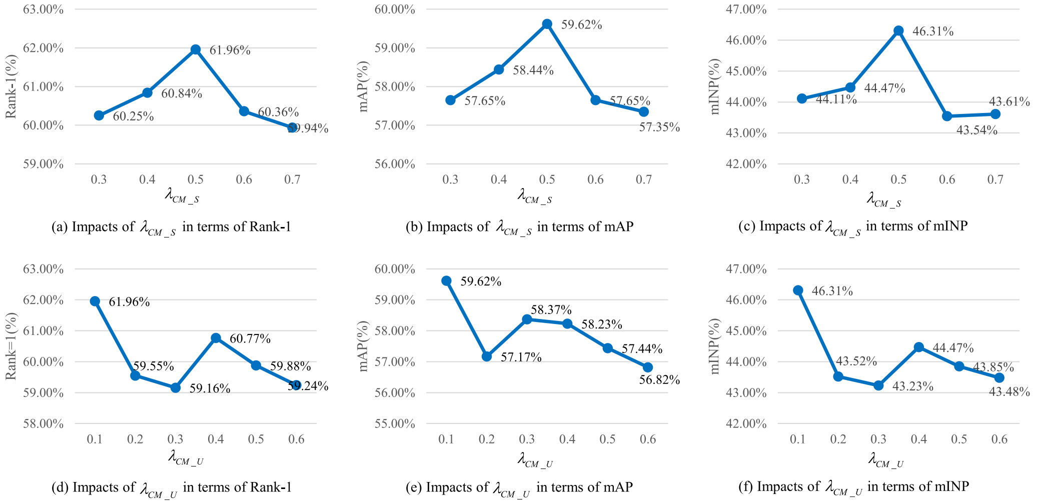
4.2 Ablation Study

Our ablation experiments are based on the implementation of our HMML framework with triplet loss. All the settings are the same except for the design of the loss function. BDTR [35] adopts a combination of L_{WM} and L_{CM_G} as its loss function, and set the tradeoff parameters as $\lambda_{WM} = 0.1$ and $\lambda_{CM\_G} = 1$. We follow these settings for these two loss terms and use them as our baseline for ablation study. Considering that the network architecture of BDTR is different from ours, we integrate its loss function into our network architecture to ensure that the performance comparison focuses on the loss modules. We conduct experiments to verify the effectiveness of the other two loss terms, L_{CM_S} and L_{CM_U}, and analyze the optimal tradeoff parameter settings for these two modules, λ_{CM_S} and λ_{CM_U}.

Table 1. Ablation Study on the RegDB and SYSU-MM01 Datasets

Datasets	SYSU-MM01					RegDB				
Settings	r=1	r=10	r=20	mAP	mINP	r=1	r=10	r=20	mAP	mINP
$L_{WM} + L_{CM\_G}$[35]	51.94	88.03	94.75	51.22	39.03	69.39	87.28	92.25	65.62	51.39
$+L_{CM\_S}$	60.99	92.03	96.76	58.54	44.85	76.46	91.58	95.02	72.48	58.25
$+L_{CM\_U}$	52.88	87.38	94.36	51.37	38.56	69.73	86.94	91.73	66.31	52.22
$+L_{CM\_S} + L_{CM\_U}$	61.96	92.51	97.07	59.62	46.31	82.97	94.03	96.42	77.56	62.75

CMC (%), mAP (%), and mINP (%).



(a) Impacts of λ_{CM_S} in terms of Rank-1

(b) Impacts of λ_{CM_S} in terms of mAP

(c) Impacts of λ_{CM_S} in terms of mINP

(d) Impacts of λ_{CM_U} in terms of Rank-1

(e) Impacts of λ_{CM_U} in terms of mAP

(f) Impacts of λ_{CM_U} in terms of mINP

Fig. 4. (a) Impacts of λ_{CM_S} in term of Rank-1(%), (b) impacts of λ_{CM_S} in term of mAP(%), and (c) impacts of λ_{CM_S} in term of mINP(%). (d) Impacts of λ_{CM_U} in term of Rank-1(%), (e) impacts of λ_{CM_U} in term of mAP(%), and (f) impacts of λ_{CM_U} in term of mINP(%).

4.2.1 *The Effectiveness of L_{CM_S} and L_{CM_U}.*

As shown in Table 1, the performance of our baseline method ($L_{WM} + L_{CM\_G}$) benefits from the better network architecture we adopted, which is considerably higher than the original result BDTR [35]. Compared to this re-implementation of BDTR, L_{CM_S} can also improve 9.05%, 7.32%, and 5.82% in terms of Rank-1, mAP, and mINP on SYSU-MM01 and 7.07%, 6.86%, and 6.86% on RegDB. However, the performance is similar when the baseline only integrates with L_{CM_U}. These results verify that the impact of modality-related variation is stronger than modality-unrelated variation in visible-infrared person Re-ID task. Although the improvement of L_{CM_U} is not that significant, the fusion of all modules can achieve the best performance. Combing all terms, HMML exceeds 10.02%, 8.4%, and 7.28% in terms of Rank-1, mAP, and mINP on SYSU-MM01 and 13.58%, 11.94%, and 11.36% on RegDB.

4.2.2 *Impacts of λ_{CM_S} and λ_{CM_U}.*

In this section, we conduct experiments to investigate the influence of λ_{CM_S} and λ_{CM_U}. For investigating the influence of λ_{CM_S}, we set $\lambda_{WM} = 0.1$, $\lambda_{CM\_G} = 1$, and $\lambda_{CM\_U} = 0.1$, and vary λ_{CM_S} from 0.3 to 0.7. The performance of different λ_{CM_S} on SYSU-MM01 are shown in Figure 4(a), (b), and (c). We can observe that about 0.5 is the optimal value of λ_{CM_S} in terms of Rank-1, mAP, and mINP. When the value of λ_{CM_S} is less than 0.5 or greater than 0.5, the performance drops. For investigating the influence of λ_{CM_U}, we set $\lambda_{WM} = 0.1$, $\lambda_{CM\_G} = 1$, and $\lambda_{CM\_S} = 0.5$ and vary λ_{CM_U} from 0.1 to 0.6. The performance of

Table 2. Generalization Ability on SYSU-MM01 Datasets

Dataset	SYSU-MM01				
Methods	r=1	r=10	r=20	mAP	mINP
Triplet Loss	51.29	88.45	95.24	50.48	37.55
HMML_T	61.96	92.51	97.07	59.62	46.31
Contrastive Loss	52.51	86.98	94.03	51.47	37.57
HMML_C	63.63	93.35	97.34	60.44	46.36

CMC (%), mAP (%), and mINP (%).

different of λ_{CM_U} on SYSU-MM01 are shown in Figure 4(d), (e), and (f). We can observe that about 0.1 is the optimal value of λ_{CM_U} in terms of Rank-1, mAP, and mINP. Although the performance has a little improvement when λ_{CM_U} is 0.3 or 0.4, the performance is the overall downward trend when the value of λ_{CM_U} is greater than 0.1.

4.3 Generalization Ability

In this section, we demonstrate the compatibility of our HMML framework with two typical pair-based metric learning, triplet loss, and contrastive loss. All settings for experiments are the same except loss implementation. HMML_T represents the implementation with triplet loss, and HMML_C represents the implementation with contrastive loss. For demonstrating that HMML is effective with both triplet loss and contrastive loss, we compare the performance with the corresponding original loss function. As shown in Table 2, HMML_T exceeds 10.67%, 9.14%, and 8.76% in terms of Rank-1, mAP, and mINP. HMML_C exceeds 11.12%, 8.97%, and 8.79% in terms of Rank-1, mAP, and mINP on SYSU-MM01. The results demonstrate that HMML is compatible with regular pair-based metric learning and can achieve outstanding performance compared with the original loss function.

4.4 Comparison with State of the Art

We compare our proposed HMML framework with state-of-the-art methods. Our approach focuses on better metric learning. Hence, the competing methods mainly involves metric learning methods for visible-infrared person Re-ID, eBDTR [35], TONE+HCML [34], HSME [13], DSCSN [36], MAC [23], TSLFN [1], HPILN [37], BDTR-AGW [22], and SIM [16]. In addition, some other outstanding methods are also compared: Zero-Padding [33], cmGAN [5], D^2RL [39], MSR [9], AlignGAN [11], Hi-CMD [4], XIV [17], cm-SSFT [21], and EDFL [19].

As shown in Table 3, HMML_C reports Rank-1 = 63.63%, mAP = 60.44, and mINP = 46.36% on SYSU-MM01 and Rank-1 = 81.82%, mAP = 76.75%, and mINP = 62.34% on RegDB. HMML_T reports Rank-1 = 61.96%, mAP = 59.62, and mINP = 46.31% on SYSU-MM01 and Rank-1 = 82.97%, mAP = 77.56%, and mINP = 62.75% on RegDB.

To compare with the best metric learning method (SIM) [16], we observe that HMML_C improves 6.58% and 6.7% in terms of Rank-1 on RegDB and SYSU-MM01 and, correspondingly, 7.73% and 5.03% improved by HMML_T. Specifically, compared to cm-SSFT [21], which shows the best performance on SYSU-MM01 by using other outstanding methods, HMML_C also can achieve 2.03% in terms of Rank-1 on SYSU-MM01 and HMML_T can achieve 0.36%. The results demonstrate the effectiveness and superiority of our proposed metric learning framework.

5 CONCLUSION

In this article, we first fully explore the pair-based similarity constraints in hybrid modality considering both class-level and modality-level. All the variations are summarized into three types:

Table 3. Comparison with State-of-the-art Methods on the RegDB and SYSU-MM01 Datasets

Datasets	RegDB			SYSU-MM01		
Methods	r=1	mAP	mINP	r=1	mAP	mINP
HOG [6]	–	–	–	3.82	2.16	–
LOMO [18]	–	–	–	1.96	1.85	–
eBDTR [35]	34.62	33.46	–	27.82	28.42	–
TONE+HCML [34]	24.44	20.08	–	14.32	16.16	–
HSME [13]	50.85	47	–	20.68	23.12	–
DSCSN [36]	60.8	60	–	35.10	37.40	–
MAC [23]	36.43	37.03	–	33.26	36.22	–
TSLFN [1]	–	–	–	56.96	54.95	–
HPILN [37]	–	–	–	41.36	42.95	–
BDTR-AGW [22]	70.05	66.37	50.19	47.5	47.65	35.30
SIM [16]	**75.24**	**78.3**	–	56.93	60.88	–
Zero-Padding [33]	17.75	18.9	–	14.8	15.95	–
cmGAN [5]	–	–	–	26.97	27.8	–
D^2RL [39]	43.4	44.1	–	28.9	29.2	–
MSR [9]	48.43	48.67	–	37.35	38.11	–
AlignGAN [11]	57.9	53.6	–	42.4	40.7	–
Hi-CMD [4]	–	–	–	34.94	35.94	–
XIV [17]	62.21	60.18	–	49.92	50.73	–
cm-SSFT [21]	–	–	–	**61.6**	**63.2**	–
EDFL [19]	49.82	51.06	–	32.91	35.17	–
HMML_C	**81.82**	**76.75**	**62.34**	**63.63**	**60.44**	**46.36**
HMML_T	**82.97**	**77.56**	**62.75**	**61.96**	**59.62**	**46.31**

Thermal images for gallery, visible images for query. Rank-1 (%), mAP(%), and mINP (%).

within-modality variation, cross-modality modality-related variation and cross-modality modality-unrelated variation. Then, we propose a comprehensive metric learning framework involving four kinds of pair-based similarity constraints to address all these variations. Based on the analysis of both class-level and modality-level similarity relationships between person images, our framework is compatible with any paired-based metric learning method. Furthermore, we give the detailed implementation of combing it with triplet loss and contrastive loss separately. Finally, extensive experiments of our approach on SYSU-MM01 and RegDB demonstrate the effectiveness and superiority of our proposed metric learning framework for visible-infrared person Re-ID.

REFERENCES

[1] Yuanxin Zhu A, Zhao Yang A, Li Wang A, Sai Zhao A, Xiao Hu A, and Dapeng Tao B. 2020. Hetero-Center loss for cross-modality person Re-identification. *Neurocomputing* 386 (2020), 97–109.

[2] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. 2017. Beyond triplet loss: A deep quadruplet network for person re-identification. In *Proceedings of the Conference on Computer Vision and Pattern Recognition.*

[3] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng. 2016. Person re-identification by multi-channel parts-based CNN with improved triplet loss function. In *Computer Vision and Pattern Recognition.*

[4] Seokeon Choi, Sumin Lee, Youngeun Kim, Taekyung Kim, and Changick Kim. 2020. 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 (CVPR'20).* IEEE.

[5] Pingyang Dai, Rongrong Ji, Haibin Wang, Qiong Wu, and Yuyu Huang. 2018. Cross-modality person re-identification with generative adversarial training. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence IJCAI-18.*

[6] N. Dalal. 2005. Histograms of oriented gradients for human detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'05)*.

[7] Nguyen Dat, Hong Hyung, Kim Ki, and Park Kang. 2017. Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors* 17, 3 (2017), 605. https://doi.org/10.3390/s17030605

[8] Weijian Deng, Liang Zheng, Guoliang Kang, Yi Yang, Qixiang Ye, and Jianbin Jiao. 2018. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE.

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

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Y. Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS'14)*. 2672–2680.

[11] Wang Guan'An, Zhang Tianzhu, Cheng Jian, Liu Si, Yang Yang, and Hou Zengguang. 2020. RGB-infrared cross-modality person re-identification via joint pixel and feature alignment. In *IEEE/CVF International Conference on Computer Vision (ICCV'19)*. IEEE.

[12] Lup Hao, Jiang Wei, Fan Xing, and Zhang Sipeng. 2019. A survey on deep learning based person re-identification. *Acta Automat. Sin.* 45, 11 (2019), 2032–2049. https://doi.org/10.16383/j.aas.c180154

[13] Yi Hao, Nannan Wang, Jie Li, and Xinbo Gao. 2019. HSME: Hypersphere manifold embedding for visible thermal person re-identification. *Proc. AAAI Conf. Artif. Intell.* 33 (2019), 8385–8392. https://doi.org/10.1609/aaai.v33i01.33018385

[14] Alexander Hermans, Lucas Beyer, and Bastian Leibe. 2017. In defense of the triplet loss for person re-identification. *CoRR* abs/1703.07737 (2017). https://doi.org/1703.07737

[15] Yan Huang, Jingsong Xu, Qiang Wu, Zhedong Zheng, Zhaoxiang Zhang, and Jian Zhang. 2018. Multi-pseudo regularized label for generated data in person re-identification. *IEEE Transactions on Image Processing* PP (2018), 1–1. https://doi.org/10.1109/TIP.2018.2874715

[16] Mengxi Jia, Yunpeng Zhai, Shijian Lu, Siwei Ma, and Jian Zhang. 2020. A similarity inference metric for RGB-infrared cross-modality person re-identification. In *Proceedings of the International Joint Conferences on Artificial Intelligence Organization*.

[17] Diangang Li, Xing Wei, Xiaopeng Hong, and Yihong Gong. 2020. Infrared-visible cross-modal person re-identification with an X modality. *Proc. AAAI Conf. Artif. Intell.* 34, 4 (2020), 4610–4617.

[18] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z. Li. 2015. Person re-identification by Local Maximal Occurrence representation and metric learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'15)*. IEEE.

[19] Haijun Liu, Jian Cheng, Wen Wang, Yanzhou Su, and Haiwei Bai. 2020. Enhancing the discriminative feature learning for visible-thermal cross-modality person re-identification. *Neurocomputing* 398 (2020), 11–19.

[20] Hao Liu, Jiashi Feng, Meibin Qi, Jianguo Jiang, and Shuicheng Yan. 2017. End-to-end comparative attention networks for person re-identification. *IEEE Trans. Image Process.* 26, 99 (2017), 3492–3506.

[21] Yan Lu, Yue Wu, Bin Liu, Tianzhu Zhang, Baopu Li, Qi Chu, and Nenghai Yu. 2020. Cross-modality person re-identification with shared-specific feature transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'20)*. IEEE.

[22] Ye Mang, Shen Jianbing, Lin Gaojie, Xiang Tao, Shao Ling, and Steven C. H. Hoi. 2021. Deep learning for person re-identification: a survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PP, 99 (2021), 1–1.

[23] Ye Mang, Lan Xiangyuan, and Leng Qingming. 2019. Modality-aware Collaborative Learning for Visible Thermal Person Re-Identification. In *27th ACM International Conference*. ACM.

[24] Xuelin Qian, Yanwei Fu, Wenxuan Wang, Tao Xiang, Yang Wu, Yu-Gang Jiang, and Xiangyang Xue. 2018. Pose-normalized image generation for person re-identification. In *Proceedings of the 15th European Conference, Munich, Germany*. Springer, Cham.

[25] Ergys Ristani and Carlo Tomasi. 2018. Features for multi-target multi-camera tracking and re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6036–6046.

[26] Hailin Shi, Yang Yang, Xiangyu Zhu, Shengcai Liao, Zhen Lei, and Stan Z. Li. 2016. Embedding deep metric for person re-identification: A study against large variations. In *European Conference on Computer Vision*. Springer, Cham, 732–748.

[27] Kihyuk Sohn. 2016. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (Eds.), Vol. 29. Curran Associates, Inc., 1857–1865.

[28] Rahul Rama Varior, Mrinal Haloi, and Gang Wang. 2016. Gated siamese convolutional neural network architecture for human re-identification. In *Proceedings of the European Conference on Computer Vision*.

[29] Rahul Rama Varior, Bing Shuai, Jiwen Lu, Dong Xu, and Gang Wang. 2016. A siamese long short-term memory architecture for human re-identification. In *Proceedings of the European Conference on Computer Vision*.

[30] Wang Guan'An, Zhang Tianzhu, Yang Yang, Cheng Jian, Chang Jianlong, Liang Xu, and Hou Zengguang. 2020. Cross-modality paired-images generation for RGB-infrared person re-identification. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 7 (2020), 12144–12151.

[31] Yicheng Wang, Zhenzhong Chen, Feng Wu, and Gang Wang. 2018. Person re-identification with cascaded pairwise convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Reconigtion.*

[32] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. 2018. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 79–88.

[33] Ancong Wu, Wei Shi Zheng, Hong Xing Yu, Shaogang Gong, and Jianhuang Lai. 2017. RGB-infrared cross-modality person re-identification. In *Proceedings of the International Conference on Computer Vision (ICCV'17).* IEEE, Los Alamitos, CA.

[34] Mang Ye, Xiangyuan Lan, Jiawei Li, and Pong Yuen. 2018. Hierarchical discriminative learning for visible thermal person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence.* AAAI.

[35] Mang Ye, Zheng Wang, Xiangyuan Lan, and Pong C. Yuen. 2018. Visible thermal person re-identification via dual-constrained top-ranking. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI'18).*

[36] Shizhou Zhang, Yifei Yang, Peng Wang, Xiuwei Zhang, and Yanning Zhang. 2021. Attend to the difference: Cross-modality person re-identification via contrastive correlation. *IEEE Transactions on Image Processing* 30 (2021), 8861–8872.

[37] Yun Bo Zhao, Jian Wu Lin, Qi Xuan, and Xugang Xi. 2020. HPILN: A feature learning framework for cross-modality person re-identification. *IET Image Process.* 13, 14 (2020), 2897–2904. https://doi.org/10.1049/iet-ipr.2019.0699

[38] Liang Zheng, Yi Yang, and Alexander G. Hauptmann. 2016. Person re-identification: Past, present and future.

[39] Wang Zhixiang, Wang Zheng, Zheng Yinqiang, Chuang Yung Yu, and Satoh Shin'Ich. 2019. Learning to reduce dual-level discrepancy for infrared-visible person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'19).*

[40] Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang. 2018. Camera style adaptation for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

Received January 2020; revised May 2021; accepted June 2021