Domain Shifting: A Generalized Solution for Heterogeneous Cross-Modality Person Re-Identification

Yan Jiang[1], Xu Cheng[1⋆], Hao Yu[2], Xingyu Liu[1], Haoyu Chen[2], and Guoying Zhao[2]

[1] School of Computer Science, Nanjing University of Information Science and Technology, Nanjing, China
[2] Center for Machine Vision and Signal Analysis, University of Oulu, Oulu, Finland
{jiangyan,xcheng,xingyu}@nuist.edu.cn
{hao.2.yu,chen.haoyu,guoying.zhao}@oulu.fi

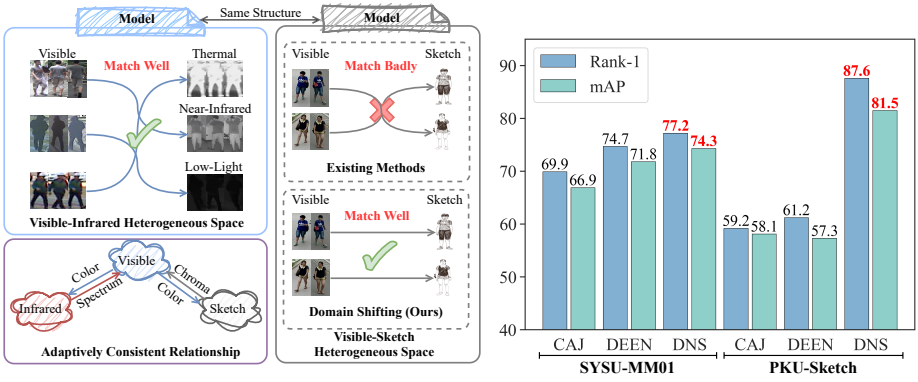
Abstract. Cross-modality person re-identification (ReID) is a challenging task that aims to match cross-modality pedestrian images across multiple camera views. Existing methods are tailored to specific tasks and perform well for visible-infrared or visible-sketch ReID. However, the performance exhibits a notable decline when the same method is utilized for multiple cross-modality ReIDs, limiting its generalization and applicability. To address this issue, we propose a generalized domain shifting method (DNS) for cross-modality ReID, which can address the generalization and perform well in both visible-infrared and visible-sketch modalities. Specifically, we propose the heterogeneous space shifting and common space shifting modules to augment specific and shared representations in heterogeneous space and common space, respectively, thereby regulating the model to learn the consistency between modalities. Further, a domain alignment loss is developed to alleviate the cross-modality discrepancies by aligning the patterns across modalities. In addition, a domain distillation loss is designed to distill identity-invariant knowledge by learning the distribution of different modalities. Extensive experiments on two cross-modality ReID tasks (*i.e.*, visible-infrared ReID, visible-sketch ReID) demonstrate that the proposed method outperforms the state-of-the-art methods by a large margin.

Keywords: Cross-modality ReID· Domain shifting · Generalization

1 Introduction

Person re-identification (ReID) aims to match the pedestrian images across multiple non-overlap cameras, which is of great practical significance for modern monitoring and criminal investigation systems [3, 16, 43, 56]. Recently, it has achieved significant progress [9,13,28,43] in addressing massive changes in viewpoint, pose, clothes, *etc*. However, the majority of previous efforts only focused

⋆ Corresponding author.



(a) Motivation of our domain shifting

(b) Performance comparison on two ReID tasks

Fig. 1: Motivation of our proposed method. CAJ and DEEN are two SOTA VI-ReID methods. SYSU-MM01 and PKU-Sketch are visible-infrared and visible-sketch ReID datasets. Our DNS can adaptively learn consistent relationships between modalities.

on visible images captured under good lighting conditions. In real-life practice, many monitoring images and videos are captured in complex conditions such as low-light or poor illumination environments. These captured low-light images and infrared images exhibit significant visual discrepancies with visible images [31, 35, 44, 46, 53]. Furthermore, when a criminal is witnessed but the surveillance camera is damaged or fails to capture the pedestrian, criminal identification can only rely on sketch images drawn by artists [1, 2, 18, 19, 22]. These situations bring challenges to current single-modality methods and raise important cross-modality ReID tasks, *e.g.*, visible-infrared, visible-sketch ReID, *etc.*

The goal of cross-modality ReID is to identify corresponding infrared, sketch, or low-light images when given a specific individual visible image, and vice versa. It is challenging due to the visual discrepancies among different modalities, as well as intra-modality variations in viewpoint, post, cloth, *etc.* [5,22,41,48,52,53]. As shown in Fig. 1 (a), visible images contain abundant color information and closely resemble the signal distribution captured by human vision. In contrast, images from other modalities often lack vital information such as color, spectrum, chroma, *etc.*, exhibiting huge discrepancies among different modalities.

To match pedestrian images across modalities, previous works [41, 43, 47, 49, 51, 53] utilize two-stream networks or Generative Adversarial Networks (GANs) [8] to mitigate the cross-modality gap in feature or image levels. Despite the encouraging progress, these works only consider specific cross-modality relationships such as color and spectrum between visible and infrared modalities well. When confronted with other paired modalities (*e.g.*, visible-sketch) with the same architecture, these approaches exhibit a notable performance degradation. As shown in Fig. 1 (b), it is obvious that CAJ [41] and DEEN [53] achieve excellent performance in visible-infrared person ReID, but they perform undesirable in visible-sketch person ReID task [22]. The reason is that these methods lack

the flexibility of learning consistent relationships between modalities, making it difficult to adapt to the differences in visible-sketch modality. This weakness significantly limits their application in real-world monitoring scenarios.

Based on the above analysis, in this paper, we propose a generalized domain shifting (DNS) approach for multiple cross-modality person ReID tasks, which solves heterogeneous cross-modality ReID tasks by considering consistent relationships in different modalities. Specifically, we introduce a distribution shifting concept and propose heterogeneous space shifting (HSS) and common space shifting (CSS), mining the consistent relationship among different modalities in heterogeneous and common space, respectively. HSS is utilized to amplify the modality-specific knowledge by shifting the distribution. Then coordinate attention (CA) [12] is exploited to strengthen the relationships of two modalities in the heterogeneous space. This process, in turn, guides the model towards focusing on modality-invariant representation, ensuring a more robust and unified understanding between different modalities. Similarly, the CSS can augment the modality-shared knowledge by shifting the distribution in common space, helping the model learn invariant representation between modalities, thereby mitigating the significant cross-modality discrepancy among modalities.

Moreover, an effective domain distillation-alignment (DDA) strategy comprising domain alignment loss ($\mathcal{L}_{da}$) and domain distillation loss ($\mathcal{L}_{dd}$) is presented to solve the cross-modality gap and identity discrepancy problems. Specifically, $\mathcal{L}_{da}$ is designed to relieve the cross-modality gap by aligning the patterns across modalities. $\mathcal{L}_{dd}$ distill the shared identity patterns by learning the distribution across modalities, keeping the identity consistency of different modalities.

Generally, our contributions are summarized as follows:

- We propose a generalized solution for heterogeneous person ReID termed domain shifting (DNS), which considers the consistent relationships among different modalities. To the best of our knowledge, it is the first work to introduce a generalized framework for multiple cross-modality ReID tasks.
- To learn the modality-invariant knowledge, we present heterogeneous space shifting (HSS) and common space shifting (CSS) by identifying consistent relationships in heterogeneous and common space, respectively.
- We carefully design the domain alignment loss and domain distillation loss to solve intra-class and inter-class discrepancies by aligning modality and distilling the shared patterns across two modalities.
- Extensive experiments on two cross-modality ReID tasks demonstrate the effectiveness and generalization of the proposed method, which outperforms the state-of-the-art approaches by a remarkable margin.

2 Related Work

2.1 Single-Modality Person ReID

Single-modality person ReID aims to match pedestrian images from multiple non-overlapping cameras, which encounters challenges such as changes in view-

point, cloth, *etc*. Existing studies [3, 40, 43] mainly focus on feature representation learning and metric learning, and achieve excellent performance in academic benchmarks. However, these works are developed for the single visible modality and can only handle intra-modality variations. In real-life scenarios, the lighting around pedestrians may change during the entire retrieval process, cameras will capture infrared images in the night or low-light scenes [31, 53]. Moreover, when surveillance cameras miss capturing specific pedestrian images, criminal recognition can only be carried out using sketch images drawn by artists [22]. These challenges raise important cross-modality ReID problems, *e.g*., visible-infrared [21, 31, 53], visible-sketch ReID [22], and are hard to handle by existing single-modality ReID methods due to the large discrepancies among modalities.

2.2 Cross-Modality Person ReID

Visible-Infrared Person ReID. Visible-infrared person re-identification (VI-ReID) is a challenging cross-modality recognition problem that aims to match visible and infrared pedestrian images across multiple camera views, as well as encountering intra-modality discrepancies such as pose variations [4,36,37,39,45]. Mainstream VI-ReID studies [41, 48, 53] focus on learning the shared representations to align visible and infrared patterns. Ye *et al*. [41] designed a channel augmentation scheme to help the two-steam model learn the shared knowledge. Yu *et al*. [48] proposed an effective task-oriented pretrained strategy to enhance the heterogeneous feature learning capacity. Zhang *et al*. [53] introduced DEEN to handle VI-ReID under low-light scenes. In addition, some generative methods [30,49,51] have been presented to generate auxiliary modalities to bridge the modality gap by Generative Adversarial Network (GAN). For instance, Wang *et al*. [30] proposed AlignGAN to translate real visible images to fake IR images for mitigating the cross-modality gap. Zhang *et al*. [51] introduced FMCNet to generate auxiliary modalities at the feature level, avoiding the introduction of interfering information. Yu *et al*. [49] designed MUN to generate robust auxiliary modality through intra-modality and cross-modality learners. However, these methods only perform well in the VI-ReID task and cannot mitigate the modality gap between visible and sketch modalities well, which limits their generalization and applicability in real-world scenes.

Visible-Sketch Person ReID. Given professional sketch queries, visible-sketch ReID (VS-ReID) aims to identify visible images from multiple cameras. It plays a key role in security systems and has achieved significant progress in recent years. Pang *et al*. [22] built the first Sketch ReID benchmark and proposed a cross-domain adversarial framework to learn domain-invariant features. Gui *et al*. [10] utilized a gradient reverse layer to process the domain discrepancies across visible and sketch. Chen *et al*. [1] designed an asymmetrical disentanglement and dynamic synthesis to explore modality-shared embedding space. Similar to VI-ReIDs, the key of VS-ReID is to learn the modality-invariant representation to mitigate the cross-modality gap. However, existing methods can hardly perform well on unfamiliar cross-modality ReID due to the lack of flexibility to learn

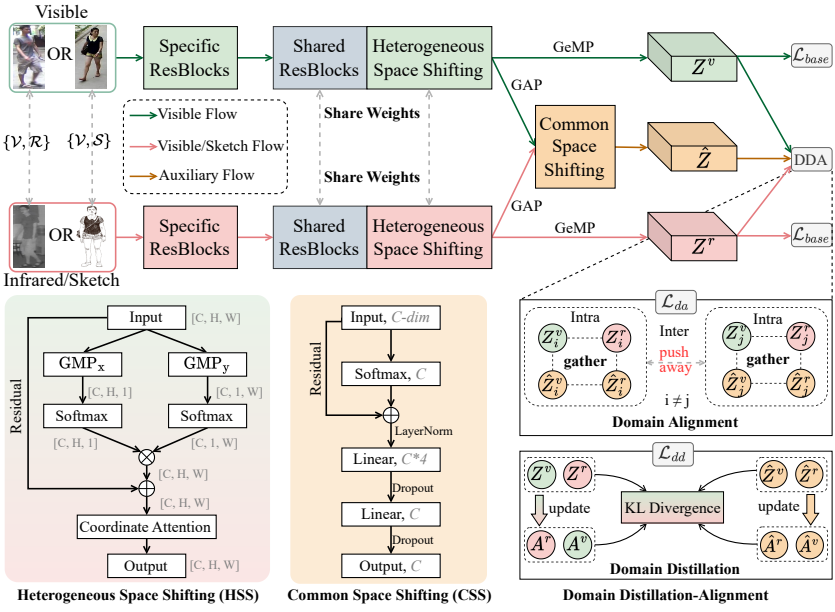


Fig. 2: The overall pipeline of the proposed Domain Shifting (DNS). GeMP and GAP are the generalized mean pooling and global average pooling, respectively. DDA denotes domain distillation-alignment strategy.

consistent relationships among heterogeneous domains. This limits their applicability in real-world surveillance systems.

To address the above-mentioned challenges, we propose domain shifting to adaptively learn modality-invariant representation among multiple modalities.

3 Methodology

In this section, we propose a novel domain shifting (DNS) method for multiple cross-modality person ReIDs. The overview of DNS is shown in Fig. 2.

3.1 Overview

Let x^m denotes the training images from the m-th modality ($m \in \{v, r, s\}$), where v, r, and s denote visible, infrared and sketch modalities, respectively. The visible, infrared, and sketch samples in the dataset are denoted as $\mathcal{V} = \{x_i^v, y_i^v\}_{i=0}^{N_v}$, $\mathcal{R} = \{x_i^r, y_i^r\}_{i=0}^{N_r}$ and $\mathcal{S} = \{x_i^s, y_i^s\}_{i=0}^{N_s}$, where N_m is the numbers of samples of each modality, and y_i^m is the corresponding identity label. $\{\mathcal{V}, \mathcal{R}\}$ and $\{\mathcal{V}, \mathcal{S}\}$ denote visible-infrared samples and visible-sketch samples, respectively. In the following subsections, we take $\{\mathcal{V}, \mathcal{R}\}$ as an example and employ $t \in \{v, r\}$ to denote the respective modalities for clear understanding.

As shown in Fig. 2, two specific ResBlocks are utilized to extract low-level modality-specific features from visible-infrared or visible-sketch images, respectively. The shared ResBlocks are exploited to learn high-semantic modality-shared patterns of two modalities. Subsequently, the heterogeneous space shifting (HSS) is designed to enhance the relationships of two modalities by augmenting modality-specific knowledge in heterogeneous space, thereby guiding the model to learn the modality-invariant representation. Furthermore, the outputs of HSS are fed into common space shifting (CSS) to learn the modality-invariant representation in the common space. Finally, DDA is developed to handle the inherent intra-class and inter-class discrepancies between modalities.

3.2 Distribution Shifting

Existing cross-modality approaches (*e.g.*, CAJ [41], DEEN [53]) are designed for specific visible-infrared ReID tasks and have achieved excellent performance. However, we found that these state-of-the-art methods are difficult to directly apply to other cross-modality person ReID tasks (*e.g.*, visible-sketch ReID). The underlying reason is that they lack the flexibility to learn consistent relationships among heterogeneous modalities, making it difficult to learn invariant knowledge in unfamiliar modalities. To overcome this issue, we introduce a shifting concept, directly shifting the domain distribution to enhance the relationships between two arbitrary modalities, thereby regulating the model to focus on learning modality-invariant knowledge. Based on shifting, we design HSS and CSS to adaptively identify consistent relationships between modalities in heterogeneous and common spaces, respectively.

Heterogeneous Space Shifting. Specifically, given the input feature maps $\mathbf{F}^t \in \mathbb{R}^{C\times H\times W}$ $(t \in \{v, r\})$ in the visible-infrared heterogeneous space, where C, H, and W denote the channel, height and width, we first utilize global max pooling to capture the most salient modality-specific knowledge in the horizontal and vertical directions, respectively. It is written as follows:

$$\mathbf{F}_h^t = GMP_x(\mathbf{F}^t), \mathbf{F}_w^t = GMP_y(\mathbf{F}^t), \quad s.t. \quad t \in \{v, r\}, \tag{1}$$

where $GMP_x(\cdot)$ and $GMP_y(\cdot)$ denote the global max pooling with pooling kernels $(h, 1)$ and $(1, w)$, respectively.

Then, the directional salient modality-specific knowledge in $\mathbf{F}_h^t$ and $\mathbf{F}_w^t$ are transformed into two distributions ranging from 0 to 1 by using a softmax function. After that, the modality-specific distribution mask is obtained by multiplying the distributions in two directions, which has the same dimension as the input $\mathbf{F}^t$. It is defined as:

$$\mathbf{M}^t = Softmax(\mathbf{F}_h^t) \otimes Softmax(\mathbf{F}_w^t), \tag{2}$$

where $\mathbf{M}^t \in \mathbb{R}^{C\times H\times W}$ is the modality-specific distribution mask. Each value in $\mathbf{M}^t$ denotes the importance of modality-specific information; $\otimes$ is the multiplication operation.

Finally, the distributions of $\mathbf{F}^t$ are shifted by residual connection to amplify the modality-specific knowledge. Further, the coordinate attention (CA) [12] is utilized to strengthen the relationships between two modalities, thereby regulating the model to focus on learning modality-invariant representation in heterogeneous space.

$$\grave{\mathbf{F}}^t = CA(\mathbf{F}^t + \mathbf{M}^t), \tag{3}$$

where $\grave{\mathbf{F}}^t \in \mathbb{R}^{C\times H\times W}$ denotes the enhanced modality-specific features.

Common Space Shifting. The CSS is presented to learn the shared representation between visible and infrared modalities in the common space. First, the enhanced modality-specific features $\grave{\mathbf{F}}^t$ processed by global average pooling (GAP) are projected into the common space by dimension reduction, which is written as:

$$\grave{\mathbf{Z}}^t = GAP(\grave{\mathbf{F}}^t), \quad s.t. \quad t \in \{v, r\}, \tag{4}$$

where $\grave{\mathbf{Z}}^t \in \mathbb{R}^{B\times C}$ is the fine-grained shared embedding; B denotes mini-batch.

Subsequently, similar to HSS, the modality-shared patterns are transformed into a distribution embedding spanning from 0 to 1 by the softmax function. After that, the distribution of $\grave{\mathbf{Z}}^t$ is shifted by residual connection to amplify the modality-shared knowledge, which is defined by:

$$\check{\mathbf{Z}}^t = \grave{\mathbf{Z}}^t + Softmax(\grave{\mathbf{Z}}^t), \tag{5}$$

where $\check{\mathbf{Z}}^t \in \mathbb{R}^{B\times C}$ is the enhanced shared patterns embedding after shifting.

Further, the modality-consistent patterns between modalities are captured within different channel dimensions by two fully connected layers (FC) in Eq. 6.

$$\hat{\mathbf{Z}}^t = Dropout(FC(Dropout(FC(LN(\check{\mathbf{Z}}^t))))), \tag{6}$$

where $\hat{\mathbf{Z}}^t \in \mathbb{R}^{B\times C}$ denotes the modality-consistent patterns. LN is the layer normalization.

In summary, HSS and CSS amplify the modality-specific and modality-shared knowledge in heterogeneous and common spaces, respectively. These two modules adaptively identify consistent relationships among different modalities, facilitating the model to learn invariant representation in two arbitrary modalities, thereby alleviating significant discrepancies in cross-modality ReID tasks.

3.3 Domain Distillation-Alignment

Domain distillation-alignment (DDA) comprises two key components: domain alignment (DA) and domain distillation (DD). DA is specifically designed to alleviate the cross-modality discrepancies, while DD focuses on distilling identity-invariant knowledge by learning the distribution in different modalities.

Domain Alignment. Following previous studies [41, 43], we first project the final features $\grave{\mathbf{F}}^t$ from the heterogeneous space into the common space using generalized mean pooling (GeMP) [25].

$$\mathbf{Z}^t = GeMP(\grave{\mathbf{F}}^t), \quad s.t. \quad t \in \{v, r\}, \tag{7}$$

where $\mathbf{Z}^t \in \mathbb{R}^{B\times C}$ is the fine-grained shared embedding in the common space. Based on $\mathbf{Z}^t$ and $\hat{\mathbf{Z}}^t$, we construct a multi-semantic representation set $\mathbf{G} = \{\mathbf{Z}^t, \hat{\mathbf{Z}}^t\}$ in the shared space, which has two advantages. (1) $\mathbf{Z}^t$ and $\hat{\mathbf{Z}}^t$ are subject to different pooling strategies, effectively capturing semantics from diverse spatial dimensions. (2) The number of samples is twice that of the original samples in the common space, which is beneficial for learning the shared relationship between the two modalities.

Subsequently, based on $\mathbf{G}$, we design an inter-class alignment loss ($\mathcal{L}_{inter}$) to mitigate the inter-class discrepancies in two modalities, which is defined as follows.

$$\mathcal{L}_{inter} = mmd(\mathbf{Z}^v, \mathbf{Z}^r) + mmd(\hat{\mathbf{Z}}^v, \hat{\mathbf{Z}}^r), \tag{8}$$

where $mmd(\cdot,\cdot)$ denotes the maximum mean discrepancy loss [14], which is defined as:

$$mmd(\mathbf{Z}^v, \mathbf{Z}^r) = ||\frac{1}{K}\sum_{i=1}^{K}\phi(\mathbf{Z}_i^v) - \frac{1}{L}\sum_{j=1}^{L}\phi(\mathbf{Z}_j^r)||_H^2, \tag{9}$$

where K and L indicate the number of samples in $\mathbf{Z}^v$ and $\mathbf{Z}^r$, respectively; $||\cdot||_H$ denotes the distribution evaluated by the Gaussian kernel function $\phi(\cdot)$, which projects features into the Hilbert space.

Furthermore, an intra-class alignment loss ($\mathcal{L}_{intra}$) is developed to solve intra-class variations, which is written as:

$$\mathcal{L}_{intra} = mmd(\mathbf{Z}^v, \hat{\mathbf{Z}}^v) + mmd(\mathbf{Z}^r, \hat{\mathbf{Z}}^r). \tag{10}$$

Finally, the domain alignment loss ($\mathcal{L}_{da}$) is defined as follows.

$$\mathcal{L}_{da} = \alpha_1\mathcal{L}_{inter} + \alpha_2\mathcal{L}_{intra}, \tag{11}$$

where α_1 and α_2 are parameters to balance the contribution of inter-class and intra-class, respectively. DA considers disparities across modalities and constrains both intra-class and inter-class variations, thus playing a pivotal role in cross-modality ReID tasks.

Domain Distillation. To alleviate the significant cross-modality gap, previous cross-modality ReID works [32–34] tend to directly utilize the Kullback Leibler (KL) divergence to regulate the distributions between modalities. However, these methods overlook the importance of preserving identity consistency between modalities, hindering cross-modality retrieval. To address this issue, we design an effective domain distillation loss ($\mathcal{L}_{dd}$) that focuses on distilling the shared identity patterns by keeping the identity consistency between modalities.

Specifically, given the fine-grained shared embeddings $\mathbf{Z}^t = \{\mathbf{z}_1^t, \mathbf{z}_2^t, ..., \mathbf{z}_B^t\}$ and $\hat{\mathbf{Z}}^t = \{\hat{\mathbf{z}}_1^t, \hat{\mathbf{z}}_2^t, ..., \hat{\mathbf{z}}_B^t\}$ in multi-semantic representation set $\mathbf{G}$, where $\mathbf{z}_i^t$ and $\hat{\mathbf{z}}_i^t$ are the i-th embeddings for two modalities in each mini-batch (B), we utilize a two-branch temporal accumulation scheme to update the shared embedding and classifier in each iteration.

Take $\mathbf{Z}^t$ as an example, the initial shared embedding and classifier can be defined as:

$$\left[\mathbf{A}^t\right]^0 = \mathbf{W}^t\left[\mathbf{Z}^t\right]^0, \mathbf{W}^t = \mathbf{C}^t, \quad s.t. \quad t \in \{v, r\}, \tag{12}$$

where $[\mathbf{A}^t]^0$ denotes the updated embedding for the t-th modality in the 0-th iteration; $\mathbf{C}^t$ represents the parameters of the classifier for the t-th modality; $\mathbf{W}^t$ is the initial classifier.

Then, the classifier is gradually updated in an accumulative manner, which is defined as follows.

$$\left[\mathbf{W}^t\right]^i = \tau * \left[\mathbf{C}^t\right]^i + (1-\tau) * \left[\mathbf{C}^t\right]^{i-1}, \tag{13}$$

where τ is the temperature parameter within the range of $(0, 1]$. Following this, the modality-invariant embedding is updated by Eq. 14:

$$[\mathbf{A}^r]^i = [\mathbf{W}^v]^i [\mathbf{Z}^r]^i, [\mathbf{A}^v]^i = [\mathbf{W}^r]^i [\mathbf{Z}^v]^i, \tag{14}$$

where $[\mathbf{Z}^t]^i = \mathbf{C}^t [\mathbf{Z}^t]^{i-1}$ denotes the i-th embedding for the t-th modality. $[\mathbf{A}^r]^i$ and $[\mathbf{A}^v]^i$ are the i-th updated infrared and visible embeddings, respectively. After that, the KL divergence and cross-entropy loss are jointly exploited to regulate the distributions of two modalities and capture the shared patterns, respectively. It is defined as follows.

$$\mathcal{L}_{dd1} = \frac{1}{k} \sum_{i=1}^{k} \left[\mathbf{A}_i^t \log \frac{\mathbf{A}_i^t}{\mathbf{Z}_i^t} - \log P(y_i^t | \mathbf{W}^t(\mathbf{Z}_i^t)) \right], \tag{15}$$

where k denotes the number of visible or infrared images in each batch; y_i^t is the *i-th* identity label. The domain-distillation loss for $\hat{\mathbf{Z}}$ is represented as $\mathcal{L}_{dd2}$, and its formulation is similar to Eq. 15 and is omitted here. The details of $\mathcal{L}_{dd2}$ are given in the supplementary material. Finally, the domain distillation loss ($\mathcal{L}_{dd}$) is summarized as follows.

$$\mathcal{L}_{dd} = \mathcal{L}_{dd1} + \mathcal{L}_{dd2}. \tag{16}$$

Overall, DDA can bridge the gap between modalities while acquiring identity-invariant patterns, which is beneficial for multiple cross-modality ReID tasks.

3.4 Final Objective Function

We employ the identity loss and the circle loss [29] as our baseline loss functions ($\mathcal{L}_{base}$). The final objective function of the proposed DNS can be written as:

$$\mathcal{L}_{total} = \mathcal{L}_{base} + \lambda \mathcal{L}_{da} + \mathcal{L}_{dd}, \tag{17}$$

where λ is the hyper-parameter.

4 Experiments

4.1 Experimental Settings

Datasets. We evaluate our proposed method in VI-ReID and VS-ReID tasks. For the VI-ReID task, we utilize the **SYSU-MM01** [31], **RegDB** [21], and

LLCM [53] datasets. The SYSU-MM01 dataset comprises 44,754 images of 491 identities captured by 4 visible and 2 infrared cameras, including 29,033 visible images and 15,712 infrared images. The RegDB dataset contains 412 identities, where each identity has 10 visible images and 10 thermal images captured by two aligned cameras. The LLCM dataset is the current large-scale dataset of images captured in both visible and infrared modes at intricate low-light scenes, including 713 identities with 25,626 visible images and 21,141 infrared images. The **PKU-Sketch** [22] dataset is utilized to evaluate our proposed method for the VS-ReID task. It contains 200 identities, where each identity has 1 sketch image drawn by artists and 2 visible images captured by cameras.

Evaluation Settings. We adopt the standard cumulative matching characteristics (CMC) and mean average precision (mAP) as the evaluation metrics.

4.2 Implementation Details

We implement our model on the Pytorch framework with one RTX 2080Ti GPU. For a fair comparison, we utilize the ResNet-50 [11] pretrained on ImageNet as our backbone, where the first two layers are adopted as specific ResBlocks, and rest layers are used as the shared ResBlocks. The proposed HSS is added after the third and fourth layers. During the training stage, all the input images are resized to 288×144 [21, 22, 31, 53]. In each mini-batch, we randomly sample 6 identities with 4 visible and 4 infrared images in the VI-ReID task, while sampling 8 identities with 1 visible and 1 sketch image in the VS-ReID task. We utilize various data augmentations, including horizontal flipping, random erasing [55], and channel exchange [41]. For VI-ReID tasks, we adopt the SGD optimizer for 100 epochs with a weight decay of $5e^{-4}$, where the initial learning rate is set to 0.2 for SYSU-MM01 and LLCM, and 0.1 for RegDB. For the VS-Sketch task, we adopt the AdamW optimizer for 100 epochs with a weight decay of 0.02, and the initial learning rate is set to 0.0009. We employ the one cycle learning rate strategy [27]. The hyper-parameters are decided by cross-validation. Specifically, we set $\alpha_1 = 0.45$, $\alpha_2 = 0.05$, $\lambda = 0.2$ and $\tau = 0.2$. Following previous methods [6, 24], the test-time augmentation is utilized during the testing stage.

4.3 Comparison with State-of-the-Art Methods

To demonstrate the superiority of our proposed DNS, we compare it with several state-of-the-art methods on the VI-ReID and VS-ReID tasks, respectively. The experimental results on the VI-ReID task are reported in Tab. 1, and the results on the VS-ReID task are reported in Tab. 2

Comparison with SOTAs on the VI-ReID task. As shown in Tab. 1, it is evident that our proposed DNS outperforms all the state-of-the-art methods on the SYSU-MM01 dataset [31]. Specifically, in the all-search mode, DNS achieves a remarkable accuracy of 77.27% for Rank-1 and 74.35% for mAP. In the indoor-search mode, DNS achieves an outstanding accuracy Rank-1 of 84.21% and an mAP of 86.83%, surpassing the second-best method by a substantial margin.

Table 1: Comparison with SOTA methods on the datasets of VI-ReID tasks. † indicates that we report the performance when no extra auxiliary information of the gallery set is introduced for a fair comparison. Rank-1 (%) and mAP (%) are reported. VIS and IR denote the visible and infrared. The red bold and blue bold front denote the best and second best performances, respectively, and the same in the following tables.

Methods	Venue	SYSU-MM01		RegDB		LLCM	
		All-Search	Indoor-Search	IR to VIS	VIS to IR	IR to VIS	VIS to IR
		Rank-1 / mAP	Rank-1 / mAP	Rank-1 / mAP	Rank-1 / mAP	Rank-1 / mAP	Rank-1 / mAP
AGW [43]	TPAMI 21	47.50 / 47.95	54.17 / 62.97	70.49 / 65.90	70.05 / 67.64	43.60 / 51.80	51.50 / 55.30
DDAG [42]	ECCV 20	54.75 / 53.02	61.02 / 67.98	68.06 / 61.80	69.34 / 63.46	40.30 / 48.40	48.00 / 52.30
LBA [23]	ICCV 21	55.41 / 54.14	66.33 / 58.46	72.43 / 65.46	74.17 / 67.64	43.80 / 53.10	50.80 / 55.60
FMCNet [51]	CVPR 22	66.34 / 62.51	68.15 / 74.09	88.38 / 83.86	89.12 / 84.43	- / -	- / -
DART [38]	CVPR 22	68.70 / 66.30	72.50 / 78.20	82.00 / 73.80	83.60 / 75.70	52.20 / 59.80	60.40 / 63.20
CAJ [41]	ICCV 21	69.88 / 66.89	76.26 / 80.37	84.55 / 77.82	85.03 / 79.14	48.80 / 56.60	56.50 / 59.80
MMN [54]	MM 21	70.60 / 66.90	76.20 / 79.60	87.50 / 80.50	91.60 / 84.10	52.50 / 58.90	59.90 / 62.70
CIFT† [17]	ECCV 22	71.77 / 67.64	78.65 / 82.11	90.12 / 84.81	92.17 / 86.96	- / -	- / -
CMT [15]	ECCV 22	71.88 / 68.57	76.90 / 79.91	91.97 / 84.46	**95.17** / 87.30	- / -	- / -
CAL [33]	ICCV 23	74.66 / 71.73	79.69 / **83.68**	**93.64** / **87.61**	94.51 / **88.67**	- / -	- / -
DEEN [53]	CVPR 23	74.70 / 71.80	**80.30** / 83.30	89.50 / 83.40	91.10 / 85.10	**54.90** / **62.90**	**62.50** / **65.80**
SGIEL [6]	CVPR 23	75.18 / 70.12	78.40 / 81.20	91.07 / 85.23	92.18 / 86.59	- / -	- / -
MUN [49]	ICCV 23	**76.24** / **73.81**	79.42 / 82.06	91.86 / 85.01	**95.19** / 87.15	- / -	- / -
DNS (Ours)	ECCV 24	**77.27** / **74.35**	**84.21** / **86.83**	**93.48** / **88.10**	93.01 / **88.56**	**57.45** / **64.11**	**66.02** / **68.60**

On the RegDB dataset [21], visible and thermal images are aligned, showing less intra-class variations. Consequently, the performance of all methods surpasses that of SYSU-MM01. In such conditions, our proposed DNS still achieves competitive results. Specifically, in IR to VIS mode, DNS attains Rank-1 of 93.48% and mAP of 88.10%. In VIS to IR mode, DNS achieves Rank-1 of 93.01% and mAP of 88.56%. Although our accuracy slightly trails behind the top-performing method, this discrepancy can be attributed to fewer intra-class variations in this dataset, limiting the full utilization of our DDA. Nonetheless, our mAP closely approaches the performance of the best method.

On the LLCM dataset [53], visible and infrared images are captured under intricate low-light conditions, giving rise to demanding intra- and inter-class variations. Consequently, the accuracy of all methods falls short of that observed in SYSU-MM01. In such challenging scenarios, our proposed DNS excels against all the SOTAs in both the IR to VIS and VIS to IR modes. This achievement can be attributed to DNS's excellent ability to obtain modality-invariant knowledge, without being affected by environmental changes. Overall, these results validate the effectiveness of our proposed DNS in the VI-ReID task.

Comparison with SOTAs on the VS-ReID task. [22]. In Tab. 2, it can be observed that our DNS delivers remarkable results, boasting an impressive 87.6% Rank-1 accuracy and 81.5% mAP accuracy on the PKU-Sketch dataset, significantly outperforming other state-of-the-art methods by a large margin. DNS can effectively learn modality-invariant knowledge due to adaptively identifying consistent relationships between two modalities, which makes it perform well not only in visible-infrared heterogeneous space but also in the visible-sketch heterogeneous space. This achievement highlights the superiority and generalization of our DNS.

Table 2: Comparison with the SOTAs on PKU-Sketch. ‡ indicates we re-implement the result with the official code.

Methods	Venue	Rank-1	Rank-5	Rank-10	mAP
TripletSN [50]	CVPR 16	9.0	26.8	42.2	-
GNSiamese [26]	TOG 16	28.9	54.0	62.4	-
AFLNet [22]	MM 18	34.0	56.3	72.5	-
CAJ‡ [41]	ICCV 21	59.2	84.4	93.4	**58.1**
CDAC [57]	TIFS 22	60.8	80.6	88.8	-
DEEN‡ [53]	CVPR 23	61.2	84.6	92.4	57.3
SketchTrans [1]	MM 22	**84.6**	**94.8**	**98.2**	-
DNS(Ours)	ECCV 24	**87.6**	**98.0**	**99.4**	**81.5**

Table 3: Ablation studies on SYSU-MM01 and PKU-Sketch.

Index	Baseline	HSS	CSS	$\mathcal{L}_{da}$	$\mathcal{L}_{kl}$	$\mathcal{L}_{dd}$	SYSU-MM01		PKU-Sketch	
							Rank-1	mAP	Rank-1	mAP
1	✓						63.21	61.69	54.2	51.7
2	✓	✓					68.71	67.05	66.8	63.2
3	✓		✓				65.47	64.39	58.6	56.5
4	✓	✓	✓				72.06	69.31	74.4	69.3
5	✓	✓	✓	✓			72.89	69.95	78.2	72.4
6	✓	✓	✓			✓	75.05	73.20	84.6	77.3
7	✓	✓	✓	✓	✓		74.65	71.41	82.4	75.7
8	✓	✓	✓	✓		✓	**77.27**	**74.35**	**87.6**	**81.5**

4.4 Ablation Study

In this subsection, we conduct ablation experiments to evaluate the effectiveness of each component introduced in this paper. The results are shown in Tab. 3.

Specifically, HSS improves the Rank-1 accuracy by 5.5% and 12.6% on the two respective datasets. Subsequently, CSS contributes to a 3.35% and 7.6% improvement in Rank-1 accuracy on the SYSU-MM01 and PKU-Sketch datasets. These notable advancements can be attributed to the outstanding performance of HSS and CSS, which enhances the modality-specific and modality-shared knowledge to strengthen consistent relationships between modalities by shifting, thereby effectively learning modality-invariant representation in heterogeneous and common space, respectively. Based on the above results, the domain alignment loss ($\mathcal{L}_{da}$) and domain distillation loss ($\mathcal{L}_{dd}$) individually improve the Rank-1 accuracy by 0.83%/3.8% and 2.99%/10.2% on two datasets. When these two loss functions work together to mitigate the inter- and intra-class discrepancies, the Rank-1 and mAP accuracy on the two datasets are significantly improved by 5.21%/5.04% and 13.2%/12.2%. By replacing $\mathcal{L}_{dd}$ with $\mathcal{L}_{kl}$, the Rank-1 accuracy and mAP accuracy on the two datasets are significantly decreased. In addition, we can find that the improvement on PKU-Sketch is significantly greater than that on SYSU-MM01. The reason is that PKU-Sketch dataset exhibits fewer

Table 4: Ablation study of the shifting in the HSS and CSS.

Method	SYSU-MM01		PKU-Sketch	
	Rank-1	mAP	Rank-1	mAP
DNS	**77.27**	**74.35**	**87.6**	**81.5**
HSS w/o shifting	73.70 ↓3.57	70.84 ↓3.51	80.0 ↓7.6	74.9 ↓6.6
CSS w/o shifting	75.16 ↓2.11	72.76 ↓1.59	81.6 ↓6.0	76.4 ↓5.1
HSS+CSS w/o shifting	72.78 ↓4.49	70.13 ↓4.22	76.2 ↓11.4	72.1 ↓9.4
CAJ + shifting	72.23 ↑2.35	70.29 ↑3.40	68.8 ↑9.6	68.4 ↑10.3
DEEN + shifting	75.99 ↑1.29	72.81 ↑1.01	70.4 ↑9.2	65.2 ↑7.9

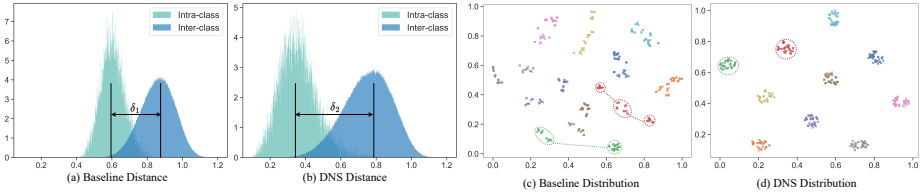


Fig. 3: Visualization of the learned features on SYSU-MM01. (a) and (b) show the intra-class and inter-class distances of cross-modality features. (c) and (d) show the distribution of feature embeddings in the 2D feature space. In the scatter charts, each color denotes an identity in the testing set. The circle and cross masks represent the features extracted from visible and infrared images.

intra-modality variations because of limited training images and the sketch images are all frontal. The major challenge lies in the significant discrepancy between visible and sketch modalities. Our DNS relieves this huge discrepancy by learning the consistent representation, resulting in greater improvement. These results demonstrate the superiority of our proposed components.

4.5 Analysis and Discussion

Intuition of Shifting. We conduct experiments to verify the effectiveness of shifting by removing the shifting of HSS and CSS. As shown in Tab. 4, by removing the shifting individually, the HSS and CSS reduce the Rank-1 accuracy by 3.57%/7.6% and 2.11%/6.0% on two datasets, respectively. When removing the shifting both in HSS and CSS, the Rank-1 accuracy are reduced by 4.49% and 11.4%. In addition, we apply the shifting to the non-local module in CAJ [41] and the MFA block in DEEN [53], which consistently improves the performance on two datasets. This demonstrates the effectiveness and generalizability of our proposed shifting in cross-modality ReID tasks.

What is DNS Doing? To elaborate on the effectiveness of DNS, we first visualize the intra- and inter-class distances of cross-modality features on SYSU-MM01. In addition, we randomly select 10 identities from SYSU-MM01 to visualize the distribution of the learned features by T-SNE [20], as shown in Fig. 3.

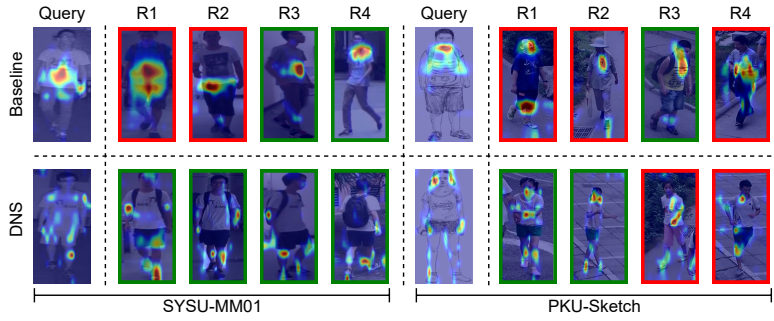


Fig. 4: Attention visualization of DNS and Baseline on SYSU-MM01 and PKU-Sketch. The red and green rectangles indicate fault and correct retrieval, respectively.

In Fig. 3 (a) and Fig. 3 (b), it is obvious that the means (the vertical lines) of inter- and intra-class distances are pushed away by DNS ($\delta_1 < \delta_2$). This means the identities in DNS are discriminated. In addition, the intra-class distance of DNS is significantly reduced compared to the baseline. This observation proves that DNS can effectively reduce the huge cross-modality gap. Meanwhile, as shown in Fig. 3 (c), the features extracted by the baseline are initially gathered into their respective centers, but the modality discrepancy remains striking (see the red and green circular dashed boxes in (c)). Moreover, the distance of different colors is not significant enough to distinguish them well. On the contrary, in Fig. 3 (d), the learned features of different identities from DNS are clustered with extremely small intra-class and giant inter-class distances. In short, DNS can effectively mitigate cross-modality discrepancy and learn modality-invariant knowledge across modalities, resulting in a remarkable improvement.

Attention Visualization. As demonstrated in Fig. 4, we conduct an infrared and sketch query to evaluate the attention maps of DNS and Baseline by XGrad-CAM [7] in different viewpoints. The attention maps obtained by baseline overlook the relevance of modalities, resulting in fallacious retrieval results. Conversely, our DNS can consistently focus on domain-invariant knowledge by identifying the consistent relationship between modalities.

5 Conclusion

This paper proposes DNS, which is the first work to handle multiple cross-modality ReID tasks. Specifically, we introduce a shifting concept, directly shifting the distribution of domains to strengthen the relationships between modalities. Following this, we propose HSS and CSS to adaptively learn consistent relationships among diverse modalities, thereby regulating the model to learn the modality-invariant knowledge between modalities. Further, DDA is designed to address intra- and inter-class discrepancies. In the future, we will explore the effectiveness of DNS to broaden the application scope in other cross-modality image recognition tasks and contribute to the community.

Acknowledgements: This research is funded in part by the Research Council of Finland (former Academy of Finland) Academy Professor project EmotionAI (Grant No. 336116, 345122), in part by the University of Oulu & Research Council of Finland Profi 7 (Grant No. 352788), in part by the National Natural Science Foundation of China (Grant No. 61802058, 61911530397), and in part by the Postgraduate Research & Practice Innovation Program of Jiangsu Province (Grant No. KYCX24_1514). We appreciate the professional and cost-effective GPU computing service provided by www.AutoDL.com.

References

1. Chen, C., Ye, M., Qi, M., Du, B.: Sketch transformer: Asymmetrical disentanglement learning from dynamic synthesis. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 4012–4020 (2022)
2. Chen, C., Ye, M., Qi, M., Du, B.: Sketchtrans: Disentangled prototype learning with transformer for sketch-photo recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
3. Chen, G., Gu, T., Lu, J., Bao, J.A., Zhou, J.: Person re-identification via attention pyramid. IEEE Transactions on Image Processing **30**, 7663–7676 (2021)
4. Cheng, X., Deng, S., Yu, H.: Exploring modality enhancement and compensation spaces for visible-infrared person re-identification. Image and Vision Computing **146**, 105040 (2024)
5. Du, Y., Zhao, Z., Su, F.: Yyds: Visible-infrared person re-identification with coarse descriptions. arXiv preprint arXiv:2403.04183 (2024)
6. Feng, J., Wu, A., Zheng, W.S.: Shape-erased feature learning for visible-infrared person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22752–22761 (2023)
7. Fu, R., Hu, Q., Dong, X., Guo, Y., Gao, Y., Li, B.: Axiom-based grad-cam: Towards accurate visualization and explanation of cnns. arXiv preprint arXiv:2008.02312 (2020)
8. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)
9. Gu, X., Chang, H., Ma, B., Bai, S., Shan, S., Chen, X.: Clothes-changing person re-identification with rgb modality only. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1060–1069 (2022)
10. Gui, S., Zhu, Y., Qin, X., Ling, X.: Learning multi-level domain invariant features for sketch re-identification. Neurocomputing **403**, 294–303 (2020)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
12. Hou, Q., Zhou, D., Feng, J.: Coordinate attention for efficient mobile network design. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13713–13722 (2021)
13. Huang, Y., Wu, Q., Xu, J., Zhong, Y., Zhang, Z.: Clothing status awareness for long-term person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11895–11904 (2021)
14. Jambigi, C., Rawal, R., Chakraborty, A.: Mmd-reid: A simple but effective solution for visible-thermal person reid. arXiv preprint arXiv:2111.05059 (2021)

15. Jiang, K., Zhang, T., Liu, X., Qian, B., Zhang, Y., Wu, F.: Cross-modality transformer for visible-infrared person re-identification. In: European Conference on Computer Vision. pp. 480–496. Springer (2022)
16. Leng, Q., Ye, M., Tian, Q.: A survey of open-world person re-identification. IEEE Transactions on Circuits and Systems for Video Technology **30**(4), 1092–1108 (2019)
17. Li, X., Lu, Y., Liu, B., Liu, Y., Yin, G., Chu, Q., Huang, J., Zhu, F., Zhao, R., Yu, N.: Counterfactual intervention feature transfer for visible-infrared person re-identification. In: European Conference on Computer Vision. pp. 381–398. Springer (2022)
18. Lin, K., Wang, Z., Wang, Z., Zheng, Y., Satoh, S.: Beyond domain gap: Exploiting subjectivity in sketch-based person retrieval. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 2078–2089 (2023)
19. Liu, X., Cheng, X., Chen, H., Yu, H., Zhao, G.: Differentiable auxiliary learning for sketch re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 3747–3755 (2024)
20. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)
21. Nguyen, D.T., Hong, H.G., Kim, K.W., Park, K.R.: Person recognition system based on a combination of body images from visible light and thermal cameras. Sensors **17**(3), 605 (2017)
22. Pang, L., Wang, Y., Song, Y.Z., Huang, T., Tian, Y.: Cross-domain adversarial feature learning for sketch re-identification. In: Proceedings of the 26th ACM international conference on Multimedia. pp. 609–617 (2018)
23. Park, H., Lee, S., Lee, J., Ham, B.: Learning by aligning: Visible-infrared person re-identification using cross-modal correspondences. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12046–12055 (2021)
24. Qiu, *et al.*: Hosnet. arXiv (2023)
25. Radenović, F., Tolias, G., Chum, O.: Fine-tuning cnn image retrieval with no human annotation. IEEE transactions on pattern analysis and machine intelligence **41**(7), 1655–1668 (2018)
26. Sangkloy, P., Burnell, N., Ham, C., Hays, J.: The sketchy database: learning to retrieve badly drawn bunnies. ACM Transactions on Graphics (TOG) **35**(4), 1–12 (2016)
27. Smith, L.N., Topin, N.: Super-convergence: Very fast training of neural networks using large learning rates. In: Artificial intelligence and machine learning for multi-domain operations applications. vol. 11006, pp. 369–386. SPIE (2019)
28. Somers, V., De Vleeschouwer, C., Alahi, A.: Body part-based representation learning for occluded person re-identification. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1613–1623 (2023)
29. Sun, Y., Cheng, C., Zhang, Y., Zhang, C., Zheng, L., Wang, Z., Wei, Y.: Circle loss: A unified perspective of pair similarity optimization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6398–6407 (2020)
30. Wang, G., Zhang, T., Cheng, J., Liu, S., Yang, Y., Hou, Z.: Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3623–3632 (2019)
31. Wu, A., Zheng, W.S., Yu, H.X., Gong, S., Lai, J.: Rgb-infrared cross-modality person re-identification. In: Proceedings of the IEEE international conference on computer vision. pp. 5380–5389 (2017)

32. Wu, J., Liu, H., Shi, W., Liu, M., Li, W.: Style-agnostic representation learning for visible-infrared person re-identification. IEEE Transactions on Multimedia (2023)
33. Wu, J., Liu, H., Su, Y., Shi, W., Tang, H.: Learning concordant attention via target-aware alignment for visible-infrared person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11122–11131 (2023)
34. Wu, Q., Dai, P., Chen, J., Lin, C.W., Wu, Y., Huang, F., Zhong, B., Ji, R.: Discover cross-modality nuances for visible-infrared person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4330–4339 (2021)
35. Wu, Z., Ye, M.: Unsupervised visible-infrared person re-identification via progressive graph matching and alternate learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9548–9558 (2023)
36. Yang, B., Chen, J., Ye, M.: Towards grand unified representation learning for unsupervised visible-infrared person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11069–11079 (2023)
37. Yang, B., Ye, M., Chen, J., Wu, Z.: Augmented dual-contrastive aggregation learning for unsupervised visible-infrared person re-identification. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 2843–2851 (2022)
38. Yang, M., Huang, Z., Hu, P., Li, T., Lv, J., Peng, X.: Learning with twin noisy labels for visible-infrared person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14308–14317 (2022)
39. Ye, M., Lan, X., Li, J., Yuen, P.: Hierarchical discriminative learning for visible thermal person re-identification. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)
40. Ye, M., Liang, C., Yu, Y., Wang, Z., Leng, Q., Xiao, C., Chen, J., Hu, R.: Person reidentification via ranking aggregation of similarity pulling and dissimilarity pushing. IEEE Transactions on Multimedia **18**(12), 2553–2566 (2016)
41. Ye, M., Ruan, W., Du, B., Shou, M.Z.: Channel augmented joint learning for visible-infrared recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13567–13576 (2021)
42. Ye, M., Shen, J., J. Crandall, D., Shao, L., Luo, J.: Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16. pp. 229–247. Springer (2020)
43. Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., Hoi, S.C.: Deep learning for person re-identification: A survey and outlook. IEEE transactions on pattern analysis and machine intelligence **44**(6), 2872–2893 (2021)
44. Ye, M., Shen, J., Shao, L.: Visible-infrared person re-identification via homogeneous augmented tri-modal learning. IEEE Transactions on Information Forensics and Security **16**, 728–739 (2020)
45. Ye, M., Wang, Z., Lan, X., Yuen, P.C.: Visible thermal person re-identification via dual-constrained top-ranking. In: IJCAI. vol. 1, p. 2 (2018)
46. Ye, M., Wu, Z., Chen, C., Du, B.: Channel augmentation for visible-infrared re-identification. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
47. Yu, H., Cheng, X., Cheng, K.H.M., Peng, W., Yu, Z., Zhao, G.: Discovering attention-guided cross-modality correlation for visible-infrared person re-identification. Pattern Recognition p. 110643 (2024)
48. Yu, H., Cheng, X., Peng, W.: Toplight: Lightweight neural networks with task-oriented pretraining for visible-infrared recognition. In: Proceedings of the

IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3541–3550 (2023)
49. Yu, H., Cheng, X., Peng, W., Liu, W., Zhao, G.: Modality unifying network for visible-infrared person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11185–11195 (2023)
50. Yu, Q., Liu, F., Song, Y.Z., Xiang, T., Hospedales, T.M., Loy, C.C.: Sketch me that shoe. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 799–807 (2016)
51. Zhang, Q., Lai, C., Liu, J., Huang, N., Han, J.: Fmcnet: Feature-level modality compensation for visible-infrared person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7349–7358 (2022)
52. Zhang, Y., Lu, Y., Yan, Y., Wang, H., Li, X.: Frequency domain nuances mining for visible-infrared person re-identification. arXiv preprint arXiv:2401.02162 (2024)
53. Zhang, Y., Wang, H.: 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. pp. 2153–2162 (2023)
54. Zhang, Y., Yan, Y., Lu, Y., Wang, H.: Towards a unified middle modality learning for visible-infrared person re-identification. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 788–796 (2021)
55. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 13001–13008 (2020)
56. Zhou, K., Yang, Y., Cavallaro, A., Xiang, T.: Learning generalisable omni-scale representations for person re-identification. IEEE transactions on pattern analysis and machine intelligence **44**(9), 5056–5069 (2021)
57. Zhu, F., Zhu, Y., Jiang, X., Ye, J.: Cross-domain attention and center loss for sketch re-identification. IEEE Transactions on Information Forensics and Security **17**, 3421–3432 (2022)