DMA: Dual Modality-Aware Alignment for Visible-Infrared Person Re-Identification

Zhenyu Cui, Jiahuan Zhou, *Member, IEEE*, and Yuxin Peng, *Senior Member, IEEE*

***Abstract*— Visible-infrared person re-identification (VI-ReID) aims to identify the same person across visible and infrared images. Its main challenge is how to extract modality-irrelevant person identity information. To alleviate cross-modality discrepancies, existing methods typically follow two paradigms: 1) Transform visible images into gray-scale color space and map them into the infrared domain. 2) Stack infrared images into RGB color space and map them into the visible domain. However, limited by different optical properties of visible and infrared waves, such mapping commonly leads to information asymmetry. Although some efforts prevent such discrepancies by data-level alignment, they typically meanwhile introduce misleading information and bring extra divergence. Therefore, existing methods fail on effectively eliminating the modality discrepancies. In this paper, we first analyze the essential factors to the generation of modality discrepancies. Secondly, we propose a novel Dual Modality-aware Alignment (DMA) model for VI-ReID, which can preserve discriminative identity information and suppress the misleading information within a uniform scheme. Particularly, based on the intrinsic optical properties of both modalities, a Dual Modality Transfer (DMT) module is proposed to perform compensation for the information asymmetry in HSV color space, thereby effectively alleviating cross-modality discrepancies and better preserving discriminative identity features. Further, an Intra-local Alignment (IA) module is proposed to suppress the misleading information, where a fine-grained local consistency objective function is designed to achieve more compact intra-class representations. Extensive experiments on several benchmark datasets demonstrate the effectiveness of our method and competitive performance with state-of-the-art methods. The source code of this paper is available at https://github.com/PKU-ICST-MIPL/DMA_TIFS2023.**

***Index Terms*— Visible-infrared person re-identification, cross-modality discrepancies, dual modality transfer, intra-local alignment.**

I. Introduction

PERSON re-identification (ReID) aims to identify the same person across different times and spaces. To extract the identity information, many researches [1], [4], [8], [9], [11] focus on deep learning-based methods [3] and have achieved great progress. However, these methods are only suitable for well-lit daytime due to the dependency on bright lighting environments. To achieve full-time ReID, recent works [17], [37] have noticed such an important but challenging ReID setting, Visible-Infrared ReID (VI-ReID), which aims to match the same person across day and night.

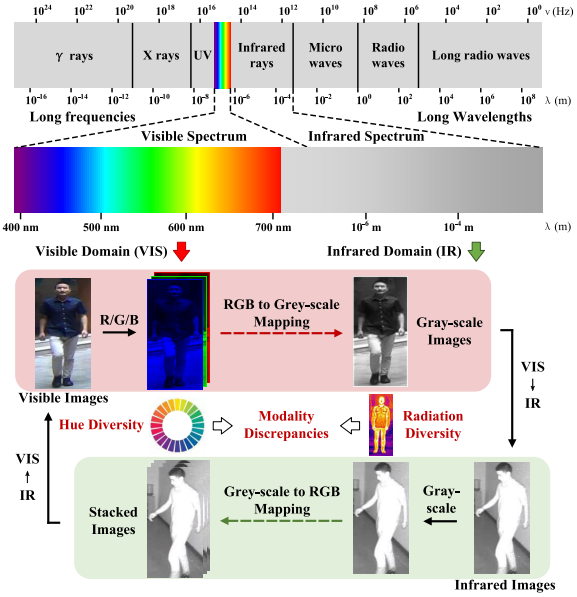


Fig. 1. The essential discrepancies between visible images and infrared images. The mapping to the RGB color space or the gray-scale color space both result in information asymmetry, remaining cross-modality discrepancies.

VI-ReID refers to identifying and matching individuals across visible and infrared images. The main challenge of VI-ReID is how to extract modality-irrelevant person identity information, resisting large modality discrepancies between the visible and infrared modalities. Inspired by general ReID methods [3], existing VI-ReID methods typically alleviate such cross-modality discrepancies based on two approaches: feature-level alignment [18], [23], [32], [58] and data-level alignment [14], [16], [51], [53].

Feature-level alignment-based methods commonly utilize neural networks to extract and map identity features from different modalities into a shared space for alignment. Nevertheless, these approaches frequently face difficulties associated with modality information, leading to a decrease in identity discrimination. Data-level alignment-based methods [13], [14], [15], [16] primarily address this challenge through two paradigms. As shown in Figure 1, some methods involve transforming visible images into gray-scale color space and mapping them into the infrared domain for representation learning. On the other hand, others forcibly stack and map

Manuscript received 13 July 2023; revised 22 September 2023 and 1 November 2023; accepted 28 November 2023. Date of publication 10 January 2024; date of current version 24 January 2024. This work was supported by the National Natural Science Foundation of China under Grant 61925201, Grant 62132001, and Grant 62376011. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Zhen Lei. *(Corresponding author: Yuxin Peng.)*

The authors are with the Wangxuan Institute of Computer Technology, and the National Key Laboratory for Multimedia Information Processing, Peking University, Beijing 100871, China (e-mail: pengyuxin@pku.edu.cn).

Digital Object Identifier 10.1109/TIFS.2024.3352408

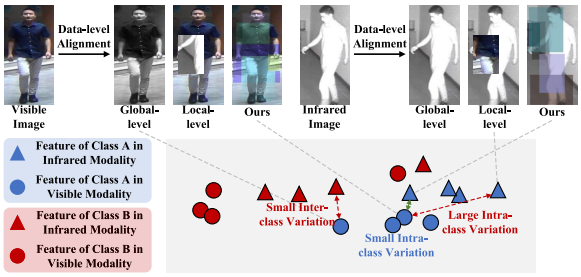


Fig. 2. Comparison of our DMA and existing methods. The above part shows the enhancement results of the existing methods (global-level [62] and local-level [65]) and our method on the visible light image and the infrared image. The following part shows the corresponding distribution of samples in the feature space.

infrared images into the RGB color space for representation learning. However, the wavelength, frequency, and radiation properties of visible waves and infrared waves are almost inconsistent. The resulting issue of inconsistent mapping exacerbates the discrepancies between modalities and prevents the extraction of the identified features. Fundamentally, the visible spectrum encompasses shorter wavelengths that allow for the perception of rich color information, whereas the infrared spectrum exhibits limited hue diversity. Additionally, infrared spectrum encompasses longer wavelengths and is sensitive to the thermal radiation emitted by the human body, whereas visible spectrum does not convey thermal radiation information. As a result, regardless of whether the mapping takes place in the RGB or gray-scale color space, current methods naturally introduce information asymmetry, thus hindering the reduction of disparities between modalities.

In order to align the visible images and infrared images in an information-symmetric domain, it is necessary to explore an appropriate color space to eliminate the cross-modal differences. However, existing alignment methods typically meanwhile introduce misleading information and bring extra divergence. As shown in Figure 2, global-level methods [51], [53], [62] often come with fuzzy edges and unrecoverable texture details, while causing radiation diversity. Besides, local-level [65], [66], [80] methods typically corrupt the most discriminative modality-common information, *e.g.*, body shapes, through randomizing image patches, which we refer to as intra-local information. As a result, these methods typically lead to large intra-class variation and small inter-class variation. Although some efforts [8], [56] alleviate the aforementioned challenges by customized constraints, the difference of intrinsic characteristics between cross-spectral features has not been explored. Therefore, existing methods typically cause misleading information dominated VI-ReID with poor performance.

We argue that by identifying a color space with jointly modelling the color information of the visible light and the radiation intensity of the infrared wave, cross-modality alignment can be facilitated by the adaptive capacity to modality discrepancies. It helps alleviate the information asymmetry across modalities, essentially contributing to the reduction of modality discrepancies. To this end, we first revisit the essence of different color spaces. HSV color space represents colors based on their hue, saturation, and value components, providing a more intuitive representation of color compared to the RGB color space. In HSV color space, the H and S channels can be employed to describe the color information of the visible light, while the V channel can be developed to characterize thermal radiation intensity. Based on the above observation, we argue that the compensation for the asymmetry of both modalities within the HSV space is the key to effectively alleviating cross-modal differences. In addition, to address the potential information corruption that arises from the data-level alignment and compensating, fine-grained local consistency constraints should be designed to improve the cross-modality intra-class consistency.

From this point, we propose a novel Dual Modality-aware Alignment (DMA) model by jointly exploring a Dual Modality Transfer (DMT) module and an Intra-local Alignment (IA) module for visible-infrared person re-identification. To align the visible images and infrared images in an information-symmetric domain, DMT module is proposed to facilitate the adaptive capacity of DMA to both modalities based on the local-level alignment. It converts visible images and infrared images to HSV color space, and then performs alignment based on the intrinsic characteristics of both modalities. As shown in Figure 2(d), our method can intuitively reduce the modality discrepancy without corrupting discriminative information. Furthermore, to suppress the misleading information, IA module introduces an intra-local center-based objective function, which effectively explores the modality-common intra-local features of both modalities. The main contributions of our work can be summarized as follows:

- We propose a novel Dual Modality-aware Alignment (DMA) model for visible-infrared person re-identification, which preserves discriminative modality-common information and suppresses the misleading information within a uniform scheme.
- A Dual Modality Transfer (DMT) module is proposed to reduce the cross-modality discrepancy in HSV color space. It effectively preserves discriminative information based on the intrinsic characteristics of both modalities, and facilitates the adaptive capacity of DMA to modality discrepancies.
- We design an Intra-local Alignment (IA) module to further suppress the misleading information, while exploring the modality-common features. A fine-grained local consistency objective function is designed to achieve more compact cross-modality intra-class representations.

The rest of this paper is organized as follows: Section II gives a brief review of related work about visible-infrared person re-identification. Section III presents the pipeline of the proposed DMA. Section IV shows the details, results, and analysis of the experiment. Section V concludes the paper.

II. Related Work

A. Visible-Infrared Person ReID (VI-ReID)

Person re-identification (Re-ID) plays an important role in security monitoring, lost person search, and intelligent system. Some methods [3], [17], [18], [56], [59] learned modality-irrelevant representations to reduce the modality discrepancy. Recently, Wu et al. [56] proposed a local pattern alignment-based pipeline to alleviate the modality discrepancy and

discover the nuances in different patterns. Zhang et al. [59] proposed a modality-specific information extraction framework to generate modality-specific features for potential key information missing. To intuitively reduce the modality discrepancy, some methods [13], [14], [15], [16] introduced GAN [61] to generate fake images at the image level. Wang et al. [14] proposed a CycleGAN-based [55] VI-ReID method to transform images into a unified domain for training and inference. Choi et al. [16] further improved the image transformation by disentangling identity-relevant and identity-irrelevant features. However, these methods are not always reliable due to the introduction of unavoidable generated noises.

B. Data-Level Alignment in VI-ReID

Data-level Alignment is a commonly-used technology in VI-ReID which can greatly improve its robustness. Some methods [51], [53], [62], [63] defined or generated a mediate modality at the global level. Ye et al. [62] proposed a channel augmentation method for VI-ReID. It randomly replaces the RGB image with a random channel to bridge both modalities. A gray-scale image-based data augmentation method is proposed in [53], which introduces a homogeneous augmented tri-modal learning method for multi-view retrieval. However, limited by the low quality and variety of the generated images, these methods generally have poor performance and efficiency. In contrast, local-level data augmentation methods enhance local discrimination. Zhong et al. [64] proposed a random erasing augmentation strategy, which facilitates local discriminative information in VI-ReID. Josi et al. [65] proposed a patch-based data augmentation method based on multi-modal image patches exchange, which improved the representation of invariant features.

However, existing data-level alignment methods mostly reduce the modality discrepancy and meanwhile suppress the most discriminative modality-common information. In this paper, we propose a Dual Modality Transfer module to perform local-level augmentation based on intrinsic characteristics of both modalities in the HSV color space and preserve the integrity of the discriminative information.

C. Color Space

To enable efficient storage and processing of images, various color spaces (i.e., RGB, YUV, and HSV) have been designed for specific applications [69]. Existing ReID methods are typically based on RGB color space [3], [70]. However, it hardly describes the influence of luminance, hue, and saturation in an intuitive way, thus suppressing the discrimination when performing data-level alignment. Therefore, some methods combine different color spaces to improve the performance of ReID systems. Tan et al. [80] focused on reducing the modality discrepancy caused by reflection coefficients of materials, and proposed a data augmentation method, called RFM, which randomizes the values in RGB channels. Nanni et al. [71] proposed an ensemble ReID system, which combines multiple color spaces to improve the robustness and effectiveness for multiple scenarios. Han et al. [72] proposed to combine histograms from multiple color spaces and improve the discriminative power for ReID systems. To overcome the clothing color over-fitting problem in VI-ReID, Zhao et al. [83] designed an HSV-based transformation method, which performs random data augmentation on visible images by parsing human bodies. However, the dependence on external knowledge and commonly designed modules limit the full development and utilization of the HSV color space.

Different from the above methods, in this paper, we propose a novel data-level alignment method, called DMA, which focuses on reducing the cross-modality discrepancy caused by the intrinsic characteristics of both visible and infrared modalities. It performs alignment in HSV color space to intuitively complement the lacking information in both modalities, thereby preserving the discrimination while reducing the modality discrepancy.

III. Dual Modality-Aware Alignment

In this section, we detail our proposed DMA method. First, the problem definition of VI-ReID is illustrated. Then, the details of the proposed DMA are presented.

A. Problem Formulation

VI-ReID aims to match the same person across different modalities. Formally, let $x^v \in \mathcal{V}$ and $x^r \in \mathcal{R}$ denote visible and infrared images respectively. $y^v \in \mathcal{Y}^v$ and $y^r \in \mathcal{Y}^r$ denote the corresponding ground truth identity labels. Given a query of a person image q^v or q^r, the goal is to match the same person based on the similarity distance between the query image q^v (q^r) and images in the gallery set $g^r(g^v)$.

The framework of the proposed Dual Modality-aware Alignment (DMA) model is shown in Figure 3, which mainly consists of two components, the Dual Modality Transfer (DMT) module and the Intra-local Alignment (IA) module. Concretely, visible images and infrared images are first fed into the DMT module to generate aligned images. Then, a single-stream backbone network is employed to generate a deep embedding for each input image. Finally, the Intra-local Alignment (IA) module is further exploited to align intra-local features.

B. Single-Stream VI-ReID Backbone

As shown in Figure 3, the backbone network of our proposed DMA is a single-stream network for both visible and infrared modalities. Given the input image x, DMA embeds x from both modalities to a same feature space and generates a deep embedding $v \in \mathbb{R}^d$ based on Global Average Pooling (GAP) and Batch Normalization (BN) [3], where d denotes the number of feature dimensions. An identification loss $\mathcal{L}_{id}$ [3] and a metric loss $\mathcal{L}_{me}$ [58], including an Euclidean constraint and a KL-divergence constraint, is employed to optimize the backbone network. By minimizing the above objective functions, DMA can preliminary preserve discriminative identity features.

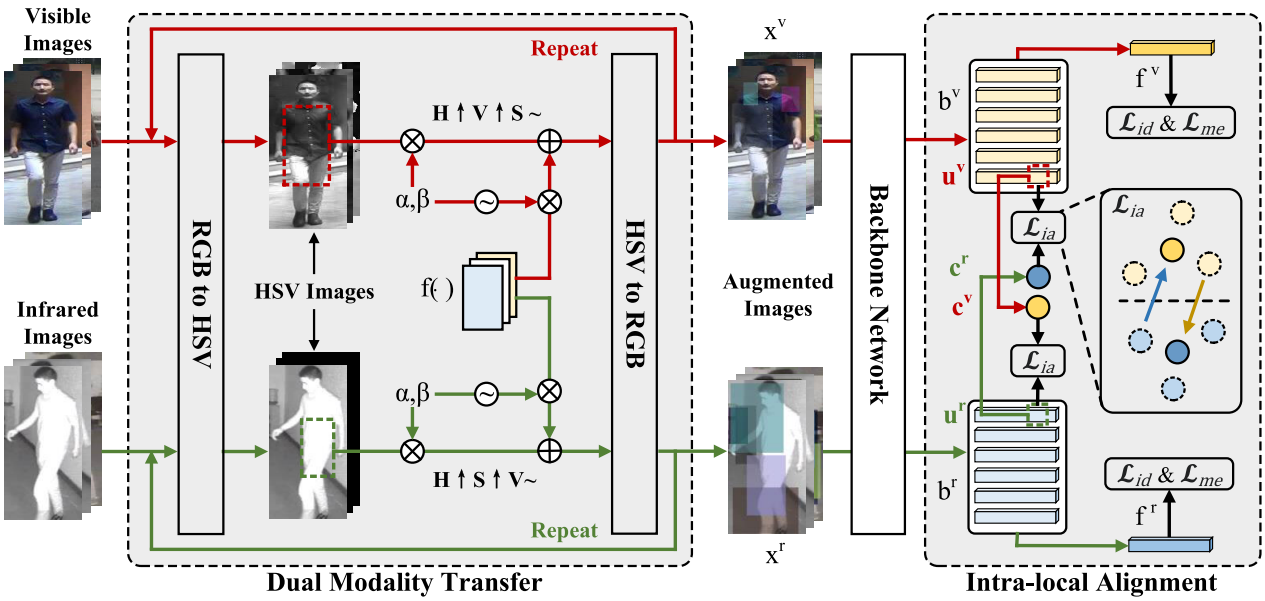


Fig. 3. The framework of our proposed Dual Modality-aware Alignment (DMA) model. The Dual Modality Transfer (DMT) module is proposed to perform local-level data alignment in HSV color space based on intrinsic characteristics of both modalities, and thus preserve the integrity of the discriminative information. Then, the Intra-local Alignment (IA) module calculates the intra-local feature-based distance of the given aligned images for exploring the modality-common features.

C. Dual Modality Transfer

The goal of the Dual Modality Transfer (DMT) module is to align images of both modalities in HSV color space based on their intrinsic characteristics and facilitate the adaptive capacity of DMA to modality changes. Specifically, visible images have a wider variety of hue and saturation, while infrared images typically have higher thermal radiation response in pedestrian-dominant regions. An intuitive solution [80] is to perform a random linear transformation on both modalities. However, directly exploiting RGB images for the above alignment is quite challenging. Therefore, DMT first transfers visible images and infrared images into the HSV color space, and performs dual modality transfer based on its intrinsic characteristics. Notably, all alignment operations in this section are normalized into $[0, 1]$ and are performed at the pixel level.

Specifically, we design a local-level data alignment strategy by directly modifying the hue (H), saturation (S) and luminance (V) maps based on image patches. The main idea is to randomly select an image patch and replace the original pixel-wise H, S, and V values by combining a directed random value in proportion. Formally, we first convert the original image of both modalities from RGB color space into HSV color space, where the RGB three channels of infrared images are stacked with the same gray-scale value. The calculation can be formulated as follows:

$$H = \frac{1}{6} \times \begin{cases} 0, & mx = mi \\ ((G-B)/\Delta) \ mod \ 6, & mx = R \\ (B-R)/\Delta + 2, & mx = G \\ (R-G)/\Delta + 4, & mx = B, \end{cases} \tag{1}$$

$$S = \begin{cases} 0 & , mx = 0 \\ \Delta / mx & , mx \neq 0, \end{cases} \tag{2}$$

$$V = mx, \tag{3}$$

where R, G, and B denote 3-channel maps in RGB color space. H, S, and V denote the corresponding map in HSV color space. mx and mi denote the maximum and minimum values in RGB color space. Δ denotes the disparity of mx and mi. Then, given the converted image t, we randomly crop and augment one of its patches $\tilde{t}$.

To reduce the modality discrepancy, DMT randomly increases the luminance of patches from visible images while randomizing the infrared ones. The transferred luminance map $V_{\tilde{t}}$ can be formulated as follows:

$$V_{\tilde{t}} = \begin{cases} (1-\alpha) \cdot V_{\tilde{t}} + \alpha \cdot f(1, 1/mx(V_{\tilde{t}})), & \tilde{t} \in \mathcal{V} \\ (1-\beta) \cdot V_{\tilde{t}} + \beta \cdot f(0, 1), & \tilde{t} \in \mathcal{R}, \end{cases} \tag{4}$$

where α denotes the random gain coefficient of the map V. β denotes the balance coefficient. $f(a, b)$ denotes a random number generator within $[a, b]$. $mx(V_{\tilde{t}})$ denotes the maximum value of V. Among them, $f(1, 1/mx(V_{\tilde{t}}))$ can promote the luminance of visible images to minimize the discrepancy of the luminance. Particularly, DMT exploits balanced randomization instead of reducing the luminance of infrared images. This is because aligning two modalities bidirectionally results in a larger modality discrepancy, and Eq. 4 avoids such potential discrepancy by effectively balancing the original image with random diversity.

Considering the lack of saturation and hue information in infrared images and the limited diversity of visible images, DMT randomizes the saturation and hue of images in both modalities. The augmented luminance map $S_{\tilde{t}}$ and the hue map $H_{\tilde{t}}$ can be formulated as follows:

$$\begin{cases} S_{\tilde{t}} = (1-\beta) \cdot S_{\tilde{t}} + \beta \cdot f(0, 1) \\ H_{\tilde{t}} = f(0, 1), \end{cases} \tag{5}$$

where β denotes the random coefficient of S, which is the same with that in Eq. 4. Finally, the transferred maps $H_{\tilde{t}}$, $S_{\tilde{t}}$

and $V_{\bar{t}}$ are merged together at the channel level. The inverse process of Eq. 1, 2 and 3 are then exploited to further generate the augmented results. Furthermore, DMT achieves diverse and robust data-level alignment by repeating the above process for 5 times.

Discussion. LTG proposed in RFM [80] is one of the most relevant methods to our proposed DMT. Compared with LTG, our proposed DMT is different from LTG in the following two main aspects: 1) The color space adopted by our proposed DMT is more reasonable and effective. The LTG module proposed in RFM uses the RGB color space to alleviate cross-modal discrepancies. However, due to the different intrinsic optical properties of both modalities, neither RGB nor gray-scale color space can effectively handle this issue due to the information asymmetry. To this end, our proposed DMA performs compensation for the information asymmetry in HSV color space, thereby effectively alleviating cross-modality discrepancies and better preserving discriminative identity features. 2) The augmentation process designed in our proposed DMT is guided by specific modalities. The augmentation strategy of LTG is a stochastic process. However, our DMT introduces a differential strategy for different modalities, which can better perform dual modality transfer based on the intrinsic characteristics of both modalities.

In summary, DMT performs dual directional data-level alignment based on the intrinsic characteristics of visible and infrared modalities. Therefore, it preserves the discrimination while intuitively reducing the cross-modality discrepancy.

D. Intra-Local Alignment

Although DMT can encourage the adaptation to cross-modality changes, it can still introduce potentially misleading intra-local information, e.g., inappropriate edges or textures, which may typically dominate VI-ReID. Therefore, the Intra-local Alignment (IA) module is proposed to exploit and align the differential intra-local information, and further improve the cross-modality intra-class consistency.

Specifically, we design an objective function to align the intra-local features. Let f^v, $f^r \in \mathbb{R}^{w\times h\times d}$ be the feature maps of visible image x^v and infrared image x^r before GAP, where w and h denote the width and the height of f^v and f^r. Inspired by [56], we split f into 6 horizontal parts (represented as N) and get separate feature blocks $f^v = [b_1^v, b_2^v, \ldots, b_N^v]$ and $f^r = [b_1^r, b_2^r, \ldots, b_N^r]$, where b_i^v and b_i^r indicate the ith part feature of the visible and infrared images. Sequentially, we further split b_i into more fine-grained parts at channel level and get L intra-local feature blocks $b_i^v = [u_{i,1}^v, u_{i,2}^v, \ldots, u_{i,L}^v]$ and $b_i^r = [u_{i,1}^r, u_{i,2}^r, \ldots, u_{i,L}^r]$. The calculation of the intra-local alignment loss $\mathcal{L}_{ia}$ can be formulated as follows:

$$\mathcal{L}_{ia} = \frac{1}{N}\cdot\frac{1}{L}\sum_{i=1}^{N}\sum_{j=1}^{L}(||u_{i,j}^r - c_{i,j}^v||_2 + ||u_{i,j}^v - c_{i,j}^r||_2), \quad (6)$$

where $c_{i,j}$ rep(resents the heterogeneous center vector of all the $u_{i,j}$ within a training batch, which has the same identity label with $u_{i,j}$.

Our proposed IA module is an improvement to the existing center-loss based objective functions [20], [23], [56], [59], [80]. It performs further subdivisions at the part-level and the channel-level, which suppresses the potentially misleading intra-local information brought by DMT, and brings more compact intra-class representations. Sequentially, it aligns cross-modal consistent information by further pulling in the cross-modal intra-local information. Benefiting from the heterogeneity of intra-class information, $\mathcal{L}_{ia}$ can suppress the misleading intra-local features that only exist in a single modality while exploring the modality-common features, thus achieving more compact intra-class representations.

E. Training and Inference

In the training phase, our proposed DMA is optimized by minimizing a multi-task loss $\mathcal{L}$, which can be formulated as follows:

$$\mathcal{L} = \mathcal{L}_{id} + \mathcal{L}_{me} + \lambda\cdot\mathcal{L}_{ia}, \quad (7)$$

where λ is a hype-parameter. $\mathcal{L}_{id}$, $\mathcal{L}_{me}$, and $\mathcal{L}_{ia}$ represent the identification loss, the metric loss, and the intra-local alignment loss.

For inference, only the embedding v is kept for VI-ReID, while DMT and IA modules are removed. Therefore, our proposed DMA does not bring any extra costs for inference.

IV. Experiments

A. Datasets and Evaluation Protocols

1) Datasets: We evaluate our method on several widely-used benchmark datasets: SYSU-MM01 [17] and RegDB [37]. SYSU-MM01 is a large-scale VI-ReID dataset, which is collected by 4 visible cameras and 2 infrared cameras. It contains 29,033 visible images and 15,712 infrared images of 491 identities. Among them, 22,258 visible images and 11,909 infrared images of 395 identities are served as training set, 3803 infrared images are served as query set, and 301/3010 randomly sampled visible images are served as gallery set for *single-shot/multi-shot* testing modes. We verify our DMA in both *all-search* and *indoor-search* scenarios with *single-shot* and *multi-shot* testing modes. RegDB contains 8,240 images of 412 identities. Among them, 206 identities with 4,120 images are served as training set, the rest are served as query set and gallery set for both *Visible to Infrared* and *Infrared to Visible* testing modes.

2) Evaluation Protocols: Following conventions in VI-ReID community, we evaluate our proposed methods with Cumulative Matching Characteristic (CMC, also denoted as R@k) curves, the mean Average Precision (mAP), and the mean Inverse Negative Penalty (mINP) [3]. We calculate R1 and mAP to comprehensively evaluate our proposed method. For SYSU-MM01, we randomly select the gallery set for 10 times to obtain stable re-identification results for a fair comparison.

B. Implementation Details

We implement our proposed DMA with PyTorch and train it on one NVIDIA A40 GPU with 48G memory. We use the commonly-used ResNet-50 [33] pretrained on ImageNet [38]

TABLE I
COMPARISON WITH STATE-OF-THE-ART METHODS ON SYSU-MM01 AND REGDB DATASETS. THE BEST RESULTS ARE BOLDED

Method	Venue	SYSU-MM01												RegDB			
		All-Search				Indoor-Search				Average				*Visible to Infrared*		*Infrared to Visible*	
		Single-Shot		*Multi-Shot*		*Single-Shot*		*Multi-Shot*		*Single-Shot*		*Multi-Shot*					
		R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP
Zero-Pad [17]	ICCV 17	14.80	15.95	19.13	10.89	20.58	26.92	24.43	18.86	17.69	21.44	21.78	14.88	-	-	-	-
cmGAN [40]	IJCAI 18	26.97	27.80	31.49	22.27	31.63	42.19	37.00	32.76	29.30	35.00	34.25	27.52	-	-	-	-
D^{2}RL [13]	CVPR 19	28.90	29.20	-	-	-	-	-	-	-	-	-	-	43.40	44.10	-	-
Hi-CMD [16]	CVPR 20	34.94	35.94	-	-	-	-	-	-	-	-	-	-	70.93	66.04	-	-
JSIA-ReID [15]	AAAI 20	38.10	36.90	45.10	29.50	43.80	52.90	52.70	42.70	40.95	44.90	48.90	36.10	48.50	49.30	48.10	48.90
AlignGAN [14]	ICCV 19	42.40	40.70	51.50	33.90	45.90	54.30	57.10	45.30	44.15	47.50	54.30	39.60	57.90	53.60	56.30	53.40
XIV [51]	AAAI 20	49.92	50.73	-	-	-	-	-	-	-	-	-	-	62.21	60.28	-	-
DDAG [25]	ECCV 20	54.75	53.02	-	-	61.02	67.98	-	-	57.89	60.50	-	-	69.34	63.46	68.06	61.80
HAT [53]	TIFS 21	55.29	53.89	-	-	62.10	69.37	-	-	58.70	61.63	-	-	71.83	67.56	70.02	66.30
NFS [54]	CVPR 21	56.91	55.45	63.51	48.56	62.79	69.79	70.03	61.45	59.85	62.62	66.77	55.00	80.54	72.10	77.95	69.79
CICL+IAMA [83]	AAAI 21	57.20	59.30	60.70	52.60	66.60	74.70	73.80	68.30	61.9	67.0	67.3	60.5	78.80	69.40	77.90	69.40
cm-SSFT [19]	CVPR 20	61.60	63.20	63.40	62.00	70.50	72.60	73.00	72.40	66.05	67.90	68.20	67.20	72.30	72.90	71.00	71.70
CM-NAS [74]	ICCV 21	61.99	60.02	68.68	53.45	67.01	72.95	76.48	65.11	64.50	66.49	72.58	59.28	84.54	80.32	82.57	78.31
MCLNet [75]	ICCV 21	65.40	61.98	-	-	72.56	76.58	-	-	68.98	69.28	-	-	80.31	73.07	75.93	69.49
SMCL [76]	ICCV 21	67.39	61.78	72.15	54.93	68.84	75.56	79.57	66.57	68.12	68.67	75.86	60.75	83.93	79.83	83.05	78.57
FMCNet [59]	CVPR 22	66.34	62.51	73.44	56.06	68.15	74.09	78.86	63.82	67.25	68.30	76.15	59.94	89.12	84.43	88.38	83.86
AGMNet [81]	JSTSP 23	69.63	66.11	-	-	74.68	78.30	-	-	72.16	72.21	-	-	88.40	81.45	85.34	81.19
MPANet [56]	CVRP 21	70.58	68.24	75.58	62.91	76.74	80.95	84.22	75.11	73.66	74.60	79.90	69.01	83.70	80.90	82.80	80.70
MAUM [60]	CVRP 22	71.68	68.79	-	-	76.97	81.94	-	-	74.33	75.37	-	-	87.87	85.09	86.95	84.34
CIFT† [58]	ECCV 22	71.77	67.64	78.00	62.46	78.65	82.11	86.97	77.03	75.21	74.88	82.49	69.75	92.17	86.96	90.12	84.81
RFM [80]	ArXiv 23	72.50	70.50	-	-	81.10	84.60	-	-	76.80	77.55	-	-	**93.50**	87.50	91.00	86.60
DEEN [82]	CVPR 23	**74.70**	**71.80**	-	-	80.30	83.30	-	-	77.50	77.55	-	-	91.10	85.10	89.50	83.40
DMA(Ours)	-	74.57	70.41	**83.03**	**65.55**	**82.85**	**85.10**	**91.33**	**80.49**	**78.71**	**77.76**	**87.18**	**73.02**	93.30	**88.34**	**91.50**	**86.80**

as our backbone network. All images are resized into 384 × 192. For the infrared image, we repeat and stack it into the 3-channel image. In the training phase, we adopt randomly flipped and erased with 50% probability for data augmentation. We use SGD to optimize the model with weight decay set to 5×10^{-4}. The initial learning rate is set to 0.01, which decays at the 80th and 140th epochs with a decay factor of 0.1. The learning rate of the pre-trained weights is set to 0.1 of the others. The coefficients α and β are set to 0.1 and 0.5, respectively. The hyper-parameter λ is set to 0.05. The batch size is set to 64 comprised of 8 identities with 4 visible images and 4 infrared images for each identity.

C. Comparison With State-of-the-Art Methods

We compare our proposed DMA with some state-of-the-art (SOTA) VI-ReID methods, including Zero-Pad [17], cmGAN [40], D^{2}RL [13], Hi-CMD [16], JSIA-ReID [15], AlignGAN [14], XIV [51], DDAG [25], HAT [53], NFS [54], CICL+IAMA [83], cm-SSFT [19], CM-NAS [74], MCLNet [75], SMCL [76], FMCNet [59], AGMNet [81], MPANet [56], MAUM [60], CIFT† [58], RFM [80], and DEEN [82]. Among them, CIFT† [58] follows the same setting as others without re-ranking.

The experimental results are shown in Table I. On SYSU-MM01 dataset, our proposed DMA outperforms all the SOTA methods on all average criteria. Our DMA achieves 74.57% R1 accuracy and 70.41% mAP accuracy in the most challenging *single-shot* all-search mode, and achieves 82.85% R1 accuracy and 85.10% mAP accuracy in the indoor-search. For the *multi-shot* mode, our DMA also superior to existing methods, especially in the indoor-search mode, which achieves 91.33% on R1 accuracy and 80.49% on mAP accuracy. The above results indicate that our DMA can effectively facilitate the retrieval of more positive people in various settings. On RegDB dataset, our DMA surpasses all SOTA methods, which achieves 93.30% and 91.50% in *Visible to Infrared* and *Infrared to Visible* settings. In summary, our DMA achieves the best mAP results (88.34% and 86.80% in *Visible to Infrared* and *Infrared to Visible* settings) and surpasses all SOTA methods.

Besides, we compared the computational consumption of our DMA with existing open-source methods on SYSU-MM01 dataset. As shown in Table III, our method requires a total of 24.3M parameters and 9.2G FLOPs. Furthermore, it can be seen that the training time and testing time of our method are moderate in both training and testing phases (taking 21h and 10s for training and inference respectively). It illustrates that our method can achieve the SOTA performance compared with existing methods at a comparable computational consumption. The above experimental results show that our DMA has a moderate computational consumption comparable to existing methods.

The superiority of our DMA can be attributed to two aspects. First, DMA effectively preserves discriminative information by performing data-level alignment in HSV color space. Second, the misleading information is further suppressed by the alignment of intra-local features. Our DMA achieves the best VI-ReID performance and reduces

TABLE II

ABLATION STUDIES ON SYSU-MM01 DATASET IN THE ALL-SEARCH MODE. THE BEST RESULTS ARE BOLDED

Baseline	DMT	IA	R1	R10	R20	mAP	mINP
✓	-	-	71.83	93.68	97.03	68.12	54.31
✓	✓	-	73.58	94.79	97.59	69.56	55.75
✓	-	✓	73.17	94.60	97.56	69.00	54.85
✓	✓	✓	**74.57**	**95.19**	**97.97**	**70.41**	**56.50**

TABLE III

COMPUTATIONAL CONSUMPTION COMPARISON ON SYSU-MM01 DATASET. THE MEMORY (PARAMS) CONSUMPTION, THE TIME (FLOPS) CONSUMPTION, THE TRAINING AND INFERENCE TIME ARE REPORTED

Method	Params	FLOPs	Training Time	Inference Time
MPANet [56]	52.5M	6.1G	4h	9s
CM-NAS [74]	24.4M	8.2G	5h	9s
DEEN [82]	89.0M	39.8G	40h	14s
DMA	**24.3M**	**9.2G**	**21h**	**10s**

TABLE IV

COMPARISON WITH EXISTING DATA-LEVEL ALIGNMENT METHODS ON SYSU-MM01 DATASET IN THE ALL-SEARCH MODE. THE BEST RESULTS ARE BOLDED

Method	R1	mAP
Baseline	71.83	68.12
Baseline+IA	73.17	69.00
Baseline+IA+CA [62]	73.95	69.40
Baseline+IA+S-PATCH [66]	72.83	68.57
Baseline+IA+M-PATCH [65]	73.52	69.49
Baseline+IA+LTG [80]	73.03	68.67
Baseline+IA+DMT(Ours)	**74.57**	**70.41**

the cross-modal discrepancies by balancing the above two advantages. Overall, the above results strongly verify the effectiveness of our proposed DMA.

D. Ablation Study

In this section, we conduct detailed ablation studies on SYSU-MM01 dataset in the all-search mode to evaluate each component of our DMA, including DMT and IA. The results are shown in Table II.

1) Baseline: We adopt the backbone network introduced in Section III-B as our baseline method.

2) Effectiveness of the Dual Modality Transfer: Compared with the baseline model, the Dual Modality Transfer module improves the R1 accuracy, mAP, and mINP by 1.75%, 1.44%, and 1.44%, respectively. It indicates that DMT module effectively suppresses the modality discrepancy and improves the VI-ReID performance. The improvement can be mainly ascribed to two reasons. For one, DMT effectively reduces the modality discrepancy by incorporating cross-modality intrinsic characteristics, and bridging both modalities in the HSV color space intuitively. For the other, DMT comprehensively preserves discriminative information while performing the above bridging, thereby encouraging the adaptation to cross-modality changes.

3) Effectiveness of the Intra-local Alignment: Compared with the baseline model, the Intra-local Alignment module brings improvements of 1.34%, 0.88%, and 0.54% on R1 accuracy, mAP accuracy, and mINP respectively. Besides, based on the baseline with DMT, our IA module brings a 0.75% improvement on mINP accuracy. This indicates that our IA module reduces the overall cost to find all the correct matches by minimizing the distance between positive samples belonging to the same person, thereby improving the intra-class consistency. The complete version of our DMA achieves the best results on SYSU-MM01 dataset in the all-search mode, achieving a gain of 2.74% and 2.29% on R1 accuracy and mAP, which verifies the effectiveness of DMA.

TABLE V

COMPARISON WITH DIFFERENT BASELINE ON SYSU-MM01 DATASET IN THE ALL-SEARCH MODE. THE BEST RESULTS ARE BOLDED

Method	R1	R10	R20	mAP	mINP
Baseline^1	62.14	89.24	94.21	60.33	47.85
Baseline^1 + DMT + IA	68.26	92.48	96.18	65.61	52.84
Baseline^2	71.83	93.68	97.03	68.12	54.31
Baseline^2 + DEEN [82]	73.29	94.58	97.49	69.27	55.50
Baseline^2 + DMT + IA	**74.57**	**95.19**	**97.97**	**70.41**	**56.50**

E. Comparison With Different Baseline Network

First, we verify the effectiveness of our proposed DMA with different baselines. Among them, Baseline^1 represents the naive baseline network reproduced by us in [80], Baseline^2 represents our baseline network introduced in Section III-B. As shown in Table V, our proposed DMA has achieved certain improvements in both two baseline networks. Specifically, DMA improves Baseline^1 by 6.12% and 5.28% on R1 and mAP accuracy, while improving Baseline^2 by 2.74% and 2.29% on R1 and mAP accuracy. It can be seen that our DMA has a greater improvement over the naive baseline. The reason is that the naive baseline lacks sufficient constraints at the feature-level, while our DMA obtains more compact intra-class representations through dual modality-aware alignment at the data-level and the feature-level, thereby significantly improving its performance. On the other hand, our method can also achieve significant improvements when using our baseline network (Baseline^2). This is mainly due to our DMA reducing the modality discrepancy through the utilization of cross-modality intrinsic characteristics. Besides, we compared DEEN [82] (one of the SOTA methods) with Baseline^2 under the same experimental setting for further analysis. It can be seen that the performance improvement of DEEN (+1.46% on R1 accuracy) is lower than that of our DMA (+2.74% on R1 accuracy). This is because DEEN did not pay attention to the modality discrepancies caused by cross-modality intrinsic characteristics, while we utilize the above characteristics to perform data-level alignment in HSV color space, thus achieving better performance. In summary, our DMA has strong generalization ability on different baselines, and achieves the SOTA performance on the baseline we introduced.

F. Comparison With Existing Data-Level Alignment Methods

We compare the proposed DMT module with several existing data-level alignment methods, including a global-level alignment method (i.e., CA [62]) and three local-level

Fig. 4. Visualization of DMT. Each column represents the results of the same person. The first row contains the input visible images and infrared images. The last row contains results generated by our DMT module. The others contain results generated by the corresponding methods.

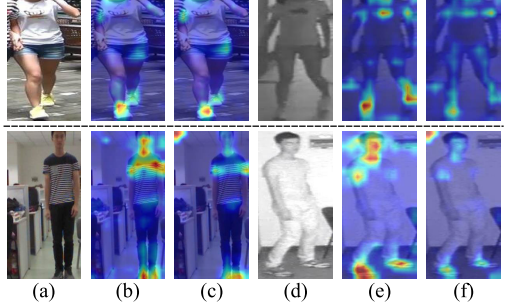


Fig. 5. Visualization of DMA. For each row, we show the results of the same person using Grad-Cam [73]. (a) and (d) are the input RGB images and the infrared images, (b) and (e) are the corresponding heatmaps generated by DMA, and (c) and (f) are those generated by the baseline method.

alignment methods (i.e., S-PATCH [66], M-PATCH [65], and LTG [80]). For a fair comparison, all the existing methods are compared under the same settings that consist of the Baseline and our proposed IA module. As shown in Table IV, our proposed DMT outperforms all the methods mentioned above, which surpasses CA [62] by 0.62% on R1 accuracy and 1.01% on mAP accuracy, and meanwhile achieves the same superiority over existing local-level alignment methods.

TABLE VI

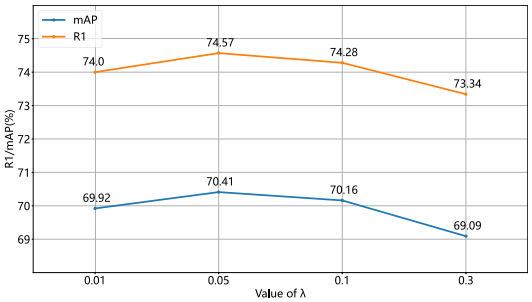
COMPARISON WITH CROSS-CENTER LOSS ON SYSU-MM01 DATASET IN THE ALL-SEARCH MODE. THE BEST RESULTS ARE BOLDED

Method	R1	mAP	mINP
Baseline	71.83	68.12	54.31
Baseline+DMT	73.58	69.56	55.75
Baseline+DMT+C	73.57	69.33	55.45
Baseline+DMT+CC	74.08	70.00	56.20
Baseline+DMT+IA (Ours)	**74.57**	**70.41**	**56.50**

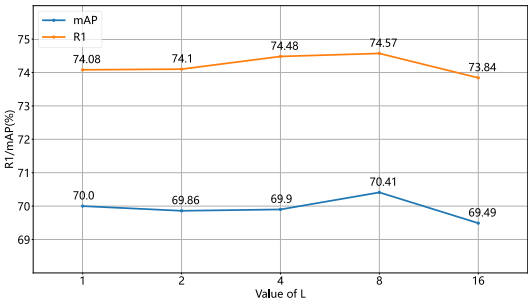
It indicates that our DMT can preserve more discriminative information, thereby achieving higher VI-ReID performance against modality discrepancies.

G. Comparison With Existing Cross-Center Loss

One of the most relevant alignment methods to our IA module is the cross-center loss proposed in [80]. To verify the effectiveness of our proposed IA module, we conduct experiments to compare the effects of the cross-center loss and IA module on SYSU-MM01 dataset. As shown in Table VI, under the same basic setting ("Baseline+DMT"), the introduction of the basic center loss (C) resulted in a certain degree of performance degradation, while the cross-center loss (CC) improves the R1 and mAP accuracy by 0.5% and 0.44%. What's more, our proposed IA module surpasses the basic setting by 0.99% and 0.85% on R1 and mAP accuracy. It indicates that the subdivisions of the feature vectors in

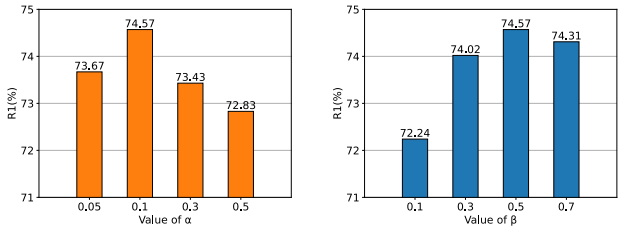


(a)



(b)

Fig. 6. The effects of hyper-parameters λ and L on SYSU-MM01 dataset in the all-search mode, which are shown in (a) and (b), respectively. R1 accuracy (%) is reported.



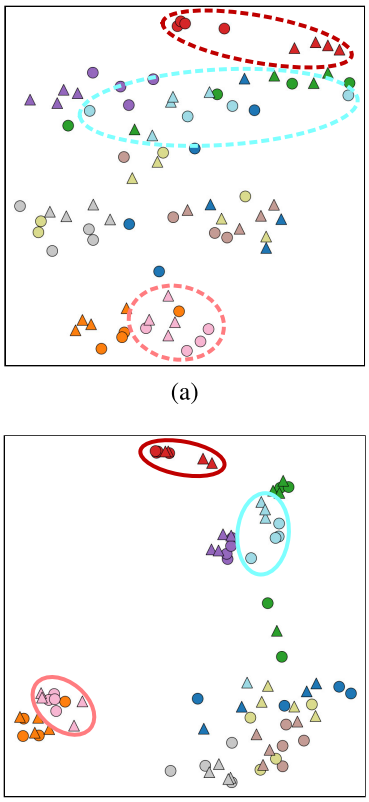
(a) (b)

Fig. 7. The effects of coefficients α and β on SYSU-MM01 dataset in the all-search mode, which are shown in (a) and (b), respectively. R1 accuracy (%) is reported.

our IA module can more effectively suppress the potentially misleading intra-local information brought by DMT. As a result, our proposed IA module brings more compact intra-class representations, which CC can hardly cover due to the lack of representation granularity. The above performance improvement mainly benefits from further subdivisions at the part-level and the channel-level, which facilitates stronger fine-grained representation and alignment capabilities. Therefore, our IA module obtains a more fine-grained intra-local representation within the local region.

H. Visualization Analysis

1) Visualization Analysis of the DMT Module: To intuitively verify the effectiveness of our Dual Modality Transfer (DMT) module, we visualize and analyse the data alignment results directly. An intuitive visualized comparison of DMT with other data-level alignment methods [62], [65], [66], [80] is shown in Figure 4. It can be seen that the results generated by CA [62] (2^{nd} row) and S-Patch [66] (3^{rd} row) can hardly eliminate cross-modality intrinsic characteristics and thus maintain the cross-modality discrepancy, while M-Patch [65] (4^{th} row)



(a) (b)

Fig. 8. The t-SNE [68] visualization of features on SYSU-MM01 dataset. The colors represent different identities, while the circles and the triangles represent visible features and infrared features, respectively. (a) and (b) are the results generated by our DMA w/o and w/ IA module.

compromises such issues by corrupting the discriminative information in the visible image, e.g. body shape, skeleton structure, etc. Moreover, LTG [80] (5^{th} row) greatly obscures the fundamental information of the image, which prevents the model from learning key discriminations that can distinguish different people. In contrast, our DMT module (the last row) can better reduce the modality discrepancy without corrupting discriminative information compared with existing methods.

Subsequently, we conducted a statistical analysis to investigate the reduction of modal differences. We randomly sampled 100 pairs of visible images and infrared images and computed their respective HSV histograms both before and after undergoing processing by the DMT module. The results are depicted in Figure 9. Upon observation, it becomes apparent that the original visible images lack hue and saturation information present in the infrared images, while the infrared images excel in capturing radiation information compared to the visible images. However, following the application of the DMT module, the aforementioned information is mutually compensated, effectively eliminating the cross-modality discrepancies. Furthermore, to verify the necessity of the retained information for VI-ReID, we apply Grad-Cam [73] to visualize discriminative regions by exploring them on the images. As shown in Figure 5, it can be observed that DMT can facilitate the perception and the alignment of discriminative information, *e.g.*, legs, shoulders, feet, etc.

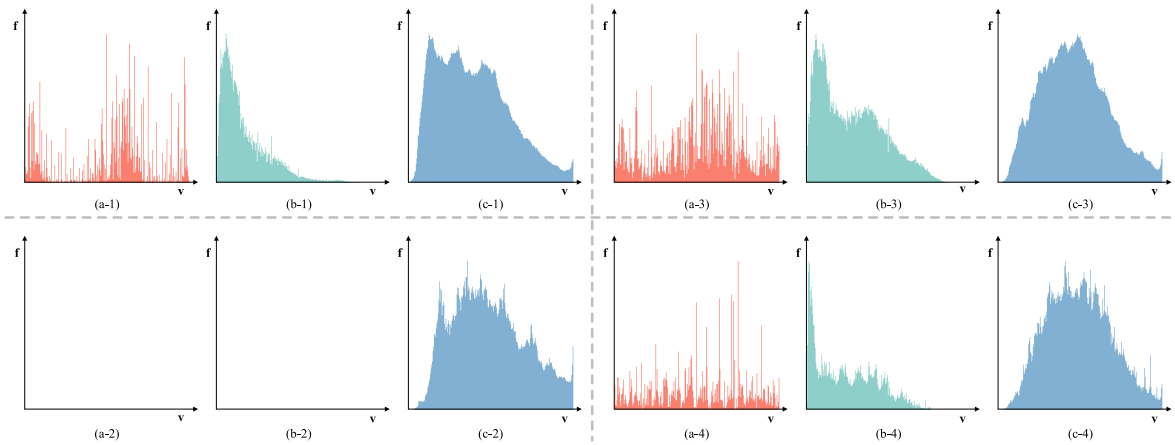


Fig. 9. The comparison of HSV histograms between original visible (infrared) images and images after using DMT. v and f represent levels of different channels and their frequencies, respectively. a, b, and c represent the histogram statistics of the three channels (H, S, and V, represented as orange, green, and blue). 1, 2, 3, and 4 represent the original visible image, the original infrared image, the visible image after using DMT and the infrared image after using DMT, respectively.

Fig. 10. Visualization of our DMA (Ours) and the baseline algorithm (B/L) on SYSU-MM01 dataset. The Rank-1 to Rank-10 results under different luminance levels are presented, where green and red boxes indicate correct and incorrect ReID results, respectively.

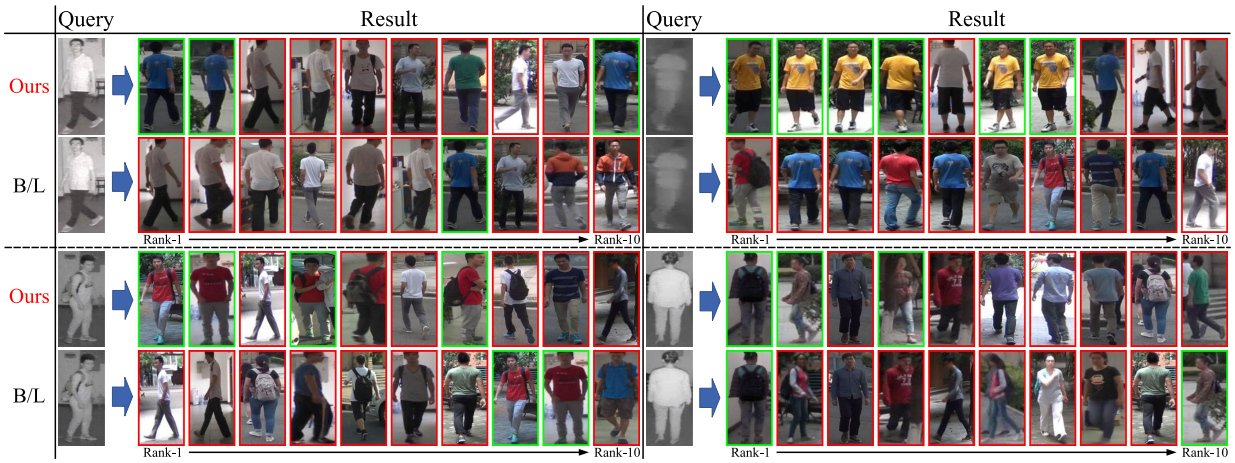


Fig. 11. Visualization of our DMA (Ours) and the baseline algorithm (B/L) on SYSU-MM01 dataset. The Rank-1 to Rank-10 results under different hues and saturations are presented, where green and red boxes indicate correct and incorrect ReID results, respectively.

2) Visualization Analysis of the IA Module: Similarly, to further verify the effectiveness of our Intra-local Alignment module, we use t-SNE [68] to visualize cross-modality features before and after using IA module. As shown in Figure 8, the features extracted by DMA after using IA are more tightly within the same class, retaining more discrimination between different classes. The above results indicate that our DMA can preserve more discriminative information, which can be further aligned by IA to achieve cross-modality alignment.

3) Visualization Analysis of the DMA Pipeline: Finally, we visualize and analyse the ReID results compared with the baseline algorithm to evaluate the overall performance of our proposed DMA model. As shown in Figure 10, compared with the baseline method, our DMA can match people more

accurately under different ambient and clothing brightness. Besides, as shown in Figure 11, our DMA helps to extract modality-irrelevant representations in the absence of color information in infrared images. The above results support that the proposed DMA can effectively reduce the cross-modality discrepancy caused by the intrinsic optical properties of both modalities

Based on the above experiments, it can be concluded that the DMT module effectively reduces the cross-modality discrepancy, while the IA module sequentially suppresses misleading intra-local information, resulting in more compact intra-class representations. These experiments further verify the effectiveness of our DMA for VI-ReID.

I. Hyper-Parameters Analysis

We first evaluate the effect of the hyper-parameter λ in Eq 7 on SYSU-MM01 dataset in the *all-search* mode. The R1 accuracy and the mAP results of DMA with different λ are shown in Figure 6 (a). The most suitable parameter setting is to set λ as 0.05, which indicates that a certain degree of the alignment of intra-local features can improve the intra-class consistency. Next, we evaluate the impact of the number of intra-local feature blocks (L). Considering that the length of the feature vector output by ResNet-50 is 2048, we select (L) as a power of 2 to achieve the average division of intra-local feature blocks. As shown in Figure 6 (b), our DMA achieves the best overall performance on R1 accuracy and mAP accuracy when L reaches 8. This indicates that an excessive or slight L will make it difficult to learn alignment information. However, our DMA achieves the best overall performance by adopting the most appropriate L. Sequentially, we compare the performance of DMA with different α and β in Eq 4 and Eq 5. As shown in Figure 7, with α increasing, the performance keeps improving before α arrives to 0.1. This is because an excessive α will corrupt the most discriminative information in RGB images. Besides, it can be observed that when β is set to 0.5, DMA achieves the best overall performance. This is because β equally balances the original image information with the alignment to a certain degree. The above results further verify the effectiveness of our method.

V. Conclusion

In this paper, we proposed a novel Dual Modality-aware Alignment (DMA) model for VI-ReID method, which is a uniform scheme to preserve discriminative information and suppress the misleading information. We first analyzed the essential factors to the generation of modality discrepancies. A Dual Modality Transfer (DMT) module is designed to perform data-level alignment in HSV color space based on the intrinsic characteristics of both modalities. It can effectively facilitate the adaptive capacity to modality discrepancies. An Intra-local Alignment (IA) module is further proposed to suppress the misleading information by introducing a fine-grained local consistency objective function, which can help to explore the modality-common intra-local features of both modalities. Extensive experimental results on several benchmark datasets demonstrate that our DMA can achieve state-of-the-art performance.

In the future, we will explore how to select or jointly utilize more suitable color spaces to implement VI-ReID across various color spaces, such as YUV, CMYK, etc., while reducing the computational consumption. Second, we will explore VI-ReID under the continual learning scenario.

References

[1] F. Wang, W. Zuo, L. Lin, D. Zhang, and L. Zhang, "Joint learning of single-image and cross-image representations for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1288–1296.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[3] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 2872–2893, Jun. 2022.

[4] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-Reciprocal encoding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3652–3661.

[5] Y. Sun, L. Zheng, W. Deng, and S. Wang, "SVDNet for pedestrian retrieval," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3820–3828.

[6] X. Qian, Y. Fu, Y.-G. Jiang, T. Xiang, and X. Xue, "Multi-scale deep learning architectures for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5409–5418.

[7] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, and Q. Tian, "Person re-identification in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3346–3355.

[8] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 480–496.

[9] Y. Suh, J. Wang, S. Tang, T. Mei, and K. M. Lee, "Part-aligned bilinear representations for person re-identification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 402–419.

[10] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based CNN with improved triplet loss function," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1335–1344.

[11] D. Li, X. Chen, Z. Zhang, and K. Huang, "Learning deep context-aware features over body and latent parts for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7398–7407.

[12] Z. Zhang, C. Lan, W. Zeng, and Z. Chen, "Densely semantically aligned person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 667–676.

[13] Z. Wang, Z. Wang, Y. Zheng, Y.-Y. Chuang, and S. Satoh, "Learning to reduce dual-level discrepancy for infrared-visible person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 618–626.

[14] G. Wang, T. Zhang, J. Cheng, S. Liu, Y. Yang, and Z. Hou, "RGB-infrared cross-modality person re-identification via joint pixel and feature alignment," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3622–3631.

[15] G.-A. Wang et al., "Cross-modality paired-images generation for RGB-infrared person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, 2020, pp. 12144–12151.

[16] S. Choi, S. Lee, Y. Kim, T. Kim, and C. Kim, "Hi-CMD: Hierarchical cross-modality disentanglement for visible-infrared person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10254–10263.

[17] A. Wu, W.-S. Zheng, H.-X. Yu, S. Gong, and J. Lai, "RGB-infrared cross-modality person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5390–5399.

[18] M. Ye, X. Lan, J. Li, and P. Yuen, "Hierarchical discriminative learning for visible thermal person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, 2018, pp. 1–9.

[19] Y. Lu et al., "Cross-modality person re-identification with shared-specific feature transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13376–13386.

[20] H. Liu, X. Tan, and X. Zhou, "Parameter sharing exploration and hetero-center triplet loss for visible-thermal person re-identification," *IEEE Trans. Multimedia*, vol. 23, pp. 4414–4425, 2021.

[21] X. Tian, Z. Zhang, S. Lin, Y. Qu, Y. Xie, and L. Ma, "Farewell to mutual information: Variational distillation for cross-modal person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1522–1531.

[22] Y. Hao, N. Wang, J. Li, and X. Gao, "HSME: Hypersphere manifold embedding for visible thermal person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 8385–8392.

[23] Y. Zhu, Z. Yang, L. Wang, S. Zhao, X. Hu, and D. Tao, "Hetero-center loss for cross-modality person re-identification," *Neurocomputing*, vol. 386, pp. 97–109, Apr. 2020.

[24] Y. Lin, A. J. Ma, and J. Wang, "Infrared-visible person re-identification via cross-modality batch normalized identity embedding and mutual learning," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2020, pp. 2296–2300.

[25] M. Ye, J. Shen, D. J Crandall, L. Shao, and J. Luo, "Dynamic dual-attentive aggregation learning for visible-infrared person re-identification," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 229–247.

[26] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li, "Learning to rank: From pairwise approach to listwise approach," in *Proc. 24th Int. Conf. Mach. Learn.*, Jun. 2007, pp. 129–136.

[27] F. Xia, T.-Y. Liu, J. Wang, W. Zhang, and H. Li, "Listwise approach to learning to rank: Theory and algorithm," in *Proc. 25th Int. Conf. Mach. Learn. (ICML)*, 2008, pp. 1192–1199.

[28] J. Revaud, J. Almazan, R. Rezende, and C. D. Souza, "Learning with average precision: Training image retrieval with a listwise loss," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5106–5115.

[29] F. Cakir, K. He, X. Xia, B. Kulis, and S. Sclaroff, "Deep metric learning to rank," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1861–1870.

[30] X. Xu, X. Yuan, Z. Wang, K. Zhang, and R. Hu, "Rank-in-rank loss for person re-identification," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 18, no. 2s, pp. 1–21, Jun. 2022.

[31] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[32] K. Jiang, T. Zhang, X. Liu, B. Qian, Y. Zhang, and F. Wu, "Cross-modality transformer for visible-infrared person re-identification," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 480–496.

[33] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[34] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "TransReID: Transformer-based object re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 14993–15002.

[35] K. Zhu et al., "AAformer: Auto-aligned transformer for person re-identification," 2021, *arXiv:2104.00921*.

[36] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1487–1495.

[37] D. Nguyen, H. Hong, K. Kim, and K. Park, "Person recognition system based on a combination of body images from visible light and thermal cameras," *Sensors*, vol. 17, no. 3, p. 605, Mar. 2017.

[38] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[39] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.

[40] P. Dai, R. Ji, H. Wang, Q. Wu, and Y. Huang, "Cross-modality person re-identification with generative adversarial training," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, p. 6.

[41] Z. Zhang, S. Jiang, C. Huang, Y. Li, and R. Y. D. Xu, "RGB-IR cross-modality person Reid based on teacher–student GAN model," *Pattern Recognit. Lett.*, vol. 150, pp. 155–161, Oct. 2021.

[42] Y. Feng, F. Chen, Y.-m. Ji, F. Wu, and J. Sun, "Efficient cross-modality graph reasoning for RGB-infrared person re-identification," *IEEE Signal Process. Lett.*, vol. 28, pp. 1425–1429, 2021.

[43] M. Ye, C. Chen, J. Shen, and L. Shao, "Dynamic tri-level relation mining with attentive graph for visible infrared re-identification," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 386–398, 2022.

[44] M. Ye, Z. Wang, X. Lan, and P. C. Yuen, "Visible thermal person re-identification via dual-constrained top-ranking," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, p. 2.

[45] K. Kansal, A. V. Subramanyam, Z. Wang, and S. Satoh, "SDL: Spectrum-disentangled representation learning for visible-infrared person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 10, pp. 3422–3432, Oct. 2020.

[46] M. Ye, X. Lan, and Q. Leng, "Modality-aware collaborative learning for visible thermal person re-identification," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 347–355.

[47] Z. Feng, J. Lai, and X. Xie, "Learning modality-specific representations for visible-infrared person re-identification," *IEEE Trans. Image Process.*, vol. 29, pp. 579–590, 2020.

[48] Y. Zhao, J. Lin, Q. Xuan, and X. Xi, "HPILN: A feature learning framework for cross-modality person re-identification," *IET Image Process.*, vol. 13, no. 14, pp. 2897–2904, Dec. 2019.

[49] A. Wu, W.-S. Zheng, S. Gong, and J. Lai, "RGB-IR person re-identification by cross-modality similarity preservation," *Int. J. Comput. Vis.*, vol. 128, no. 6, pp. 1765–1785, Jun. 2020.

[50] Y. Hao, N. Wang, X. Gao, J. Li, and X. Wang, "Dual-alignment feature embedding for cross-modality person re-identification," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 57–65.

[51] D. Li, X. Wei, X. Hong, and Y. Gong, "Infrared-visible cross-modal person re-identification with an X modality," in *Proc. AAAI*, vol. 34, no. 4, 2020, pp. 4610–4617.

[52] Y. Ling, Z. Zhong, Z. Luo, P. Rota, S. Li, and N. Sebe, "Class-aware modality mix and center-guided metric learning for visible-thermal person re-identification," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 889–897.

[53] M. Ye, J. Shen, and L. Shao, "Visible-infrared person re-identification via homogeneous augmented tri-modal learning," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 728–739, 2021.

[54] Y. Chen, L. Wan, Z. Li, Q. Jing, and Z. Sun, "Neural feature search for RGB-infrared person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 587–597.

[55] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2242–2251.

[56] Q. Wu et al., "Discover cross-modality nuances for visible-infrared person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4328–4337.

[57] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 499–515.

[58] X. Li et al., "Counterfactual intervention feature transfer for visible-infrared person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, Tel Aviv, Israel. Cham, Switzerland: Springer, Oct. 2022, pp. 381–398.

[59] Q. Zhang, C. Lai, J. Liu, N. Huang, and J. Han, "FMCNet: Feature-level modality compensation for visible-infrared person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 7339–7348.

[60] J. Liu, Y. Sun, F. Zhu, H. Pei, Y. Yang, and W. Li, "Learning memory-augmented unidirectional metrics for cross-modality person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 19344–19353.

[61] I. Goodfellow et al., "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[62] M. Ye, W. Ruan, B. Du, and M. Z. Shou, "Channel augmented joint learning for visible-infrared recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 13547–13556.

[63] Z. Huang, J. Liu, L. Li, K. Zheng, and Z.-J. Zha, "Modality-adaptive mixup and invariant decomposition for RGB-infrared person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 1, 2022, pp. 1034–1042.

[64] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, 2020, pp. 13001–13008.

[65] A. Josi, M. Alehdaghi, R. M. O. Cruz, and E. Granger, "Multimodal data augmentation for visual-infrared person Reid with corrupted data," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. Workshops (WACVW)*, Jan. 2023, pp. 1–10.

[66] M. Chen, Z. Wang, and F. Zheng, "Benchmarks for corruption invariant person re-identification," 2021, *arXiv:2111.00880*.

[67] B. Farou, H. Rouabhia, H. Seridi, and H. Akdag, "Novel approach for detection and removal of moving cast shadows based on RGB, HSV and YUV color spaces," *Comput. Informat.*, vol. 36, no. 4, pp. 837–856, 2017.

[68] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 1–27, 2008.
[69] P. Kay, B. Berlin, L. Maffi, W. R. Merrifield, and R. Cook, *The World Color Survey*. Princeton, NJ, USA: Citeseer, 2009.
[70] Q. Leng, M. Ye, and Q. Tian, "A survey of open-world person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 4, pp. 1092–1108, Apr. 2020.
[71] L. Nanni, M. Munaro, S. Ghidoni, E. Menegatti, and S. Brahnam, "Ensemble of different approaches for a reliable person re-identification system," *Appl. Comput. Informat.*, vol. 12, no. 2, pp. 142–153, Jul. 2016.
[72] K. Han, W. Wan, G. Chen, and L. Hou, "Person re-identification using multiple features fusion," in *Proc. Int. Conf. Audio, Lang. Image Process. (ICALIP)*, Jul. 2016, pp. 409–413.
[73] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.
[74] C. Fu, Y. Hu, X. Wu, H. Shi, T. Mei, and R. He, "CM-NAS: Cross-modality neural architecture search for visible-infrared person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 11803–11812.
[75] X. Hao, S. Zhao, M. Ye, and J. Shen, "Cross-modality person re-identification via modality confusion and center aggregation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 16383–16392.
[76] Z. Wei, X. Yang, N. Wang, and X. Gao, "Syncretic modality collaborative learning for visible infrared person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 225–234.
[77] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Pose-driven deep convolutional model for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3980–3989.
[78] L. Zheng, Y. Huang, H. Lu, and Y. Yang, "Pose-invariant embedding for deep person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4500–4509, Sep. 2019.
[79] Y.-J. Cho and K.-J. Yoon, "Improving person re-identification via pose-aware multi-shot matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1354–1362.
[80] L. Tan et al., "Exploring invariant representation for visible-infrared person re-identification," 2023, *arXiv:2302.00884*.
[81] H. Liu, D. Xia, and W. Jiang, "Towards homogeneous modality learning and multi-granularity information exploration for visible-infrared person re-identification," *IEEE J. Sel. Topics Signal Process.*, vol. 17, no. 3, pp. 1–15, Jan. 2023.
[82] Y. Zhang and H. Wang, "Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 2153–2162.
[83] Z. Zhao, B. Liu, Q. Chu, Y. Lu, and N. Yu, "Joint color-irrelevant consistency learning and identity-aware modality adaptation for visible-infrared cross modality person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 4, May 2021, pp. 3520–3528.

Zhenyu Cui received the B.S. degree in computer science and technology from the China University of Petroleum (East China), Qingdao, China, in 2018, and the M.S. degree in computer science from the University of Chinese Academy of Sciences, Beijing, China, in 2021. He is currently pursuing the Ph.D. degree with the Wangxuan Institute of Computer Technology, Peking University. His current research interests include computer vision and deep learning.

Jiahuan Zhou (Member, IEEE) received the B.E. degree from Tsinghua University in 2013 and the Ph.D. degree from the Department of Electrical Engineering and Computer Science, Northwestern University, in 2018. During Summer 2018, he was a Research Intern with Microsoft Research, Redmond, WA, USA. From 2019 to 2022, he was a Post-Doctoral Fellow and a Research Assistant Professor with Northwestern University. Currently, he is a Tenure-Track Assistant Professor with the Wangxuan Institute of Computer Technology, Peking University. He has authored more than 20 papers in international journals and conferences, including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON IMAGE PROCESSING, CVPR, ICCV, AAAI, and ECCV. His current research interests include computer vision, deep learning, and machine learning. He serves as the Area Chair for CVPR, ICME, ICPR; an Associate Editor for *Journal of Machine Vision and Applications* (Springer); and a regular Reviewer Member for several journals and conferences, such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, *International Journal of Computer Vision*, IEEE TRANSACTIONS ON IMAGE PROCESSING, CVPR, ICCV, ECCV, NeurIPS, and ICML.

Yuxin Peng (Senior Member, IEEE) received the Ph.D. degree in computer application technology from Peking University, Beijing, China, in 2003. He is currently the Boya Distinguished Professor with the Wangxuan Institute of Computer Technology and the National Key Laboratory for Multimedia Information Processing, Peking University. He has authored over 200 papers, including more than 100 papers in top-tier journals and conference proceedings. He has submitted 51 patent applications and has been granted 39 of them. His current research interests include cross-media analysis and reasoning, image and video recognition and understanding, and computer vision. He led his team to win First Place in video semantic search evaluation of TRECVID ten times in recent years. He won the First Prize of the Beijing Science and Technology Award in 2016 (ranking first) and the First Prize of the Scientific and Technological Progress Award of the Chinese Institute of Electronics in 2020 (ranking first). He was a recipient of the National Science Fund for Distinguished Young Scholars of China in 2019 and the Best Paper Award at MMM 2019 and NCIG 2018. He serves as an Associate Editor for IEEE TRANSACTIONS ON MULTIMEDIA and IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY.