Prototype-Based Diversity and Integrity Learning for All-Day Multi-Modal Person Re-Identification

Zi Wang, Chenglong Li, *Member, IEEE*, Pengyu Li, Aihua Zheng, Jin Tang, and Bin Luo

***Abstract*—Recent multi-modal person re-identification methods have improved model performance by leveraging complementary information from multiple spectra. However, existing methods cannot ensure feature stability under varying illumination and rely on inflexible paired data, remaining inadequate against real-world cross-time retrieval and modality-missing challenges. To solve these, we first propose diversity representation that augments illumination-sensitive images to simulate diverse lighting conditions via illumination augmentation and enriches instance features using modality-specific prototypes via multiple interaction modules. Secondly, we propose integrity reconstruction that leverages prototypes and available instance features to recover information, the reconstruction module effectively utilizes identity and modality cues to address unpredictable missing problems. In addition, we build a more comprehensive dataset (AllDay843) to alleviate the inadequate dataset diversity, which comprises 91,371 images of 843 identities captured by multi-modal cameras across various periods throughout the day, while incorporating numerous real-world challenges. By integrating diversity representation and integrity reconstruction, the proposed Prototype-Based Diversity and Integrity learning network (PDINet) establishes excellence on the AllDay843 dataset, surpassing existing state-of-the-art approaches. The data and codes are available in GitHub.**

***Index Terms*—Person re-identification, multi-modal learning, prototype learning, cross-time retrieval, incomplete modality.**

I. Introduction

OVER the last decade, there has been an exponential increase in single-modal and cross-modal person re-identification (ReID) datasets and methods [1], [2], [3], [4], [5], [6], [7], [8], [9]. Recently, to mitigate the impact of low light and nighttime scenes, researchers have developed several multi-modal object ReID datasets that incorporate complementary information from visible (RGB), near-infrared (NIR), and thermal infrared (TIR) data sources [10], [11], [12]. By exploiting the complementary nature of multi-modal information, existing approaches have substantially improved performance on complex tasks compared to conventional single-modal and cross-modal methods. Nevertheless, practical real-world scenarios present distinctive challenges that can lead to suboptimal retrieval accuracy when approaches focus excessively on multi-modal learning.

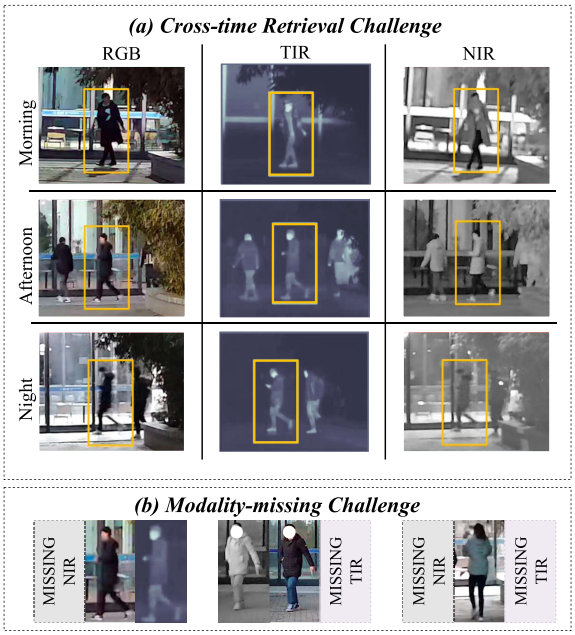


Fig. 1. (a) Cross-time retrieval challenge. Three cameras capture images of the same person (in yellow boxes) at different times. (b) Modality-missing challenge. Several modalities of missingness may occur during the testing phase, such as missing images in NIR or TIR.

First, the same pedestrian may appear at different time periods, posing the challenge of cross-time retrieval. According to the first column of Fig. 1 (a), RGB images captured by optical sensors can record the color and texture information of objects under favorable lighting conditions. TIR sensors offer more discriminative information for distinguishing between human bodies and the surrounding environment or additional attire, as illustrated in the second column. NIR modality can overcome issues of low illumination, as depicted in the third column, providing additional information, especially during afternoon periods. However, current multi-modal collaborative learning approaches, including both complementary infor-

Received 1 May 2025; revised 12 August 2025; accepted 17 September 2025. Date of publication 29 September 2025; date of current version 29 October 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62372003 and Grant 62376004; in part by the Natural Science Foundation of Anhui Province under Grant 2308085Y40; in part by the Project of Key Laboratory of Intelligent Computing and Signal Processing (Anhui University), Ministry of Education under Grant 2024A004; and in part by the Ministry of Education Key Laboratory of Brain-inspired Intelligent Perception and Cognition, University of Science and Technology of China under Grant 2421004. The associate editor coordinating the review of this article and approving it for publication was Dr. Jie Wen. *(Corresponding author: Aihua Zheng.)*

Zi Wang is with the School of Computer Science and Technology, Anhui University, Hefei 230093, China, also with the Key Laboratory of Intelligent Computing and Signal Processing, Ministry of Education, Anhui University, Hefei 230093, China, and also with the Ministry of Education Key Laboratory of Brain-inspired Intelligent Perception and Cognition, University of Science and Technology of China, Hefei 230026, China.

Chenglong Li, Pengyu Li, and Aihua Zheng are with the Information Materials and Intelligent Sensing Laboratory of Anhui Province, Hefei 230093, China, also with the Anhui Provincial Key Laboratory of Intelligent Detection and Diagnosis for Traffic Infrastructure, Hefei 230093, China, and also with the School of Artificial Intelligence, Anhui University, Hefei 230093, China (email: ahzheng214@foxmail.com).

Jin Tang and Bin Luo are with the School of Computer Science and Technology, Anhui University, Hefei 230093, China.

Digital Object Identifier 10.1109/TIFS.2025.3615708

TABLE I
DETAILED SETTINGS OF THE PROPOSED ALLDAY843 DATASET AND MULTIPLE DERIVED DATASETS

Image Number	**AllDay843**			**AllDay843-G**			**AllDay843-C (RN)**			**AllDay843-C (RT)**			**AllDay843-C (NT)**		
	Train	Query	Gallery	Train	Query	Gallery	Train	Query	Gallery	Train	Query	Gallery	Train	Query	Gallery
Visible	23482	4685	4655	23482	4685	4655	23482	4685	-	23482	4685	-	-	-	-
Infrared	23482	3127	3093	23482	4685	4655	23482	-	4685	-	-	-	23482	4685	-
Thermal	23482	2689	2676	23482	4685	4655	-	-	-	23482	-	4655	23482	-	4655
Total	70446	10501	10424	70446	14055	13965	46964	4685	4685	46964	4685	4655	46964	4685	4655

TABLE II
COMPARISON OF ALLDAY843, ALLDAY843-G, ALLDAY843-C WITH RELATED DATASETS

	Datasets	Typical Challenges				Cross-time Retrieval	Modality Missing	Identity Number	Modality			
		High Illu.	Low Illu.	Low Res.	Thermal Cross				RGB	NI	TI	others
Single Modal	CUHK03 [52]	-	-	-	-	-	-	1467	13164	-	-	-
	Market1501 [1]	-	-	✓	-	-	-	1501	32668	-	-	-
	MSMT17 [53]	-	-	✓	-	-	-	4101	126441	-	-	-
	MARS [54]	-	-	-	-	-	-	1261	1191003	-	-	-
Cross Modal	SYSU-MM01 [2]	-	-	✓	-	✓	-	491	287628	15792	-	-
	RegDB [55]	-	-	-	✓	-	-	412	4120	-	4120	-
	LLCM [5]	-	✓	-	-	-	-	1064	25626	21141	-	-
	PKU Sketch [56]	-	-	-	-	-	-	200	400	-	-	200
	AllDay843-C (RN)	✓	✓	✓	✓	✓	-	**843**	**28167**	**28137**	-	-
	AllDay843-C (RT)	✓	✓	✓	✓	✓	-	**843**	**28167**	-	**28137**	-
	AllDay843-C (NT)	✓	✓	✓	✓	✓	-	**843**	-	**28167**	**28137**	-
Multi Modal	BIWI [57]	-	-	-	-	-	-	78	-	-	-	-
	PAVIS [58]	-	-	-	-	-	-	79	395	-	-	395
	3DPes [59]	-	-	-	-	-	-	200	1012	-	-	1012
	CAVIAR4REID [60]	-	-	-	-	-	-	72	1220	-	-	1220
	RGBNT201 [10]	✓	✓	✓	✓	✓	-	201	4787	4787	4787	-
	AllDay843	✓	✓	✓	✓	✓	✓	**843**	**32822**	**29702**	**28847**	-
	AllDay843-G	✓	✓	✓	✓	✓	✓	**843**	**32822**	**32822**	**32822**	-

mation fusion [10], [13], [14] and modality-specific feature learning [15], [16], [17], [18], are susceptible to illumination-sensitive data variations, resulting in compromised training stability and hindered acquisition of diverse representations.To address the challenges posed by cross-time retrieval, we propose diversity representation. The illumination augmentation augments RGB images to simulate diverse lighting conditions (e.g., morning, evening) based on time labels, improving illumination robustness. Learnable modality-specific prototypes interact with instance features extracted from all modalities, embedding domain-specific information through the interaction module. And the subspace constraint loss ensures that augmented features remain distinctive and inter-modality features remain consistent. By enforcing orthogonality within the augmented features and minimizing multi-modal feature distances, this loss enhances both the richness and stability of the features.

Secondly, data loss or corruption caused by multi-device collaboration often leads to modality-missing challenge, which is common in real-world scenarios, as illustrated in Fig. 1 (b). Some approaches [10], [13], [18] have explored feature transformation techniques to address data deficiencies, thereby enhancing model adaptability. However, accurate pedestrian representations consist of both modality-related and identity-related cues, making it difficult for feature transformation techniques to recover complete information fully. To address the challenges posed by modality missing, we propose integrity reconstruction. For samples with missing modalities (e.g., NIR), the reconstruction module restores the absent information by analyzing the relationship between available RGB instance features and the NIR prototype. The missing modality data is recovered uniquely for each sample, aligning with the corresponding identity and modality characteristics. Additionally, the subspace constraint loss also supports effective information reconstruction.

Thirdly, existing multi-modal person ReID datasets (e.g., [10]) are limited in scale and collected exclusively during nighttime, which constrains the generalization ability of trained models. To address this limitation, we assemble data from real-world challenging environments under diverse illumination conditions to create the AllDay843 dataset and define the new evaluation protocol for the all-day person ReID task by filtering out images captured from the same period. The AllDay843 dataset comprises 91,371 multi-modal images of 201 individuals captured by six non-overlapping cameras, offering richer complementary information and a larger number of identities. It includes images recorded at different times (morning, afternoon, evening) and under various weather conditions (sunny, cloudy, rainy, foggy, snowy). Furthermore, AllDay843 incorporates complex challenges such as cross-temporal retrieval and modality-missing challenges, enhancing its relevance to real-world scenarios and providing a valuable

resource for exploring the application of multi-modal ReID technologies. More details and comparisons are shown in Table II.

By integrating diversity representation and completeness reconstruction, the proposed Prototype-based Diversity and Integrity Learning Network (PDINet) effectively mitigates both challenges of cross-time retrieval and modality-missing. Through collecting multi-modal data under realistic and challenging conditions, AllDay843 exhibits closer alignment with real-world scenarios compared to existing datasets. This work represents the pioneering effort in introducing the task of all-day multi-modal person re-identification, along with its corresponding backbone and benchmark dataset. The main contributions of this paper can be summarized as follows:

- We propose the Prototype-based Diversity and Integrity learning network (PDINet) tailored for the all-day multi-modal person ReID task. The diversity representation focuses on augmenting RGB images across lighting conditions and enhancing features with modality-specific prototypes. The integrity reconstruction centers around recovering missing modality data using prototype-guided relationships. The subspace constraint loss also ensures augmented feature distinction and multi-modal feature consistency.
- We establish a comprehensive benchmark dataset called AllDay843 that more closely aligns with real-world challenges. It consists of 91,371 images of 843 different identities recorded by 6 non-overlapping cameras. AllDay843 contains pedestrian images captured several times throughout the day, including the morning, afternoon, and evening. Moreover, AllDay843 involves the modality-missing issue that several images are absent in the testing phase.
- We demonstrate the importance of multi-modal information in solving all-day person ReID tasks through extensive and sufficient comparative experiments. Meanwhile, we compare the proposed PDINet with the state-of-the-art methods on the AllDay843 dataset to demonstrate its superiority. Ablation experiments demonstrate the respective effectiveness of components.

II. Related Work

In this section, we first review related works on complementary information fusion and modality-specific feature learning in multi-modal person ReID. Then, we introduce several multi-modal object ReID datasets that provide diverse challenges for the research community.

A. Multi-Modal Person Re-Identification Methods

Early approaches based on convolutional neural networks dominated the research landscape. PFNet [10] employs parameter-independent multi-branch networks to extract spectral features, partitions global features into multiple parts, and performs hierarchical fusion at both global and local levels. Wang et al. [15] propose an interaction, embedding, enhancement framework that enabled stage-aware learning of modality-specific information during training. The advent of Transformer [19] architectures has opened new opportunities for performance improvements in multi-modal person ReID. Wang et al. [13] introduce an object-centric selection approach, utilizing spatial-frequency selection modules to filter background interference and aggregating cross-spectral features into unified representations. Wang et al. [16] propose a heterogeneous testing-retraining framework that improved descriptor discriminability by constraining inter-spectral sample distribution distances, leading to better generalization on unseen data. Zheng et al. [20] develop a glare-aware cross-spectral enhancement network that adaptively restores glare-corrupted features under the guidance of thermal spectra. Wang et al. [18] addresses dynamic quality variations in multi-modal imaging by proposing a novel feature learning framework that adaptively balances decoupled features through Mixture-of-Experts architecture. Recently, vision-language foundation models have been adapted for ReID tasks. MambaPro [14] achieved efficient object ReID through parallel adapter tuning and cross-spectral prompt transformation, effectively integrating complementary information with low computational overhead. IDEA [21] leverages the multi-modal large language model to perform structured and concise text annotation, while integrating multi-modal information with text-derived semantic guidance, thus generating more robust features in complex scenarios. However, current approaches for either complementary fusion or modality-aware learning are sensitive to illumination variations, adversely affecting both training stability and representation diversity.

B. Multi-Modal Object Re-Identification Datasets

To cope with complex illumination changes in real scenes, infrared modalities that are complementary to RGB modalities and are stable to illumination are introduced in some computer vision tasks, such as RGB-thermal tracking [22], [23], [24], [25], [26], RGB-thermal saliency detection [27], [28], [29], [30], [31], etc. In recent years, researchers have begun exploring the application value of multi-modal data in object ReID to overcome the performance limitations of visible light imaging devices in nighttime environments. Against this backdrop, pioneering multi-modal vehicle ReID and person ReID datasets [10], [11] have emerged, innovatively integrating three modalities: RGB (visible light), NIR (near-infrared), and TIR (thermal infrared). Their primary contribution lies in establishing the first benchmark platform for multi-modal collaboration, systematically validating the complementary advantages of multi-source data under low-light conditions. As research progressed, to address the practical challenge of cross-view vehicle retrieval, Zheng et al. [12] propose the MSVR310 dataset, which collects multi-view synchronous observation data, annotates precise multi-view vehicle information, and specifically defines evaluation protocols for cross-view retrieval. Notably, the strong glare interference from nighttime vehicle lighting systems can significantly degrade imaging quality. To tackle this, the WMVeID863 [20] dataset was introduced, focusing on capturing data under extreme lighting conditions with glare

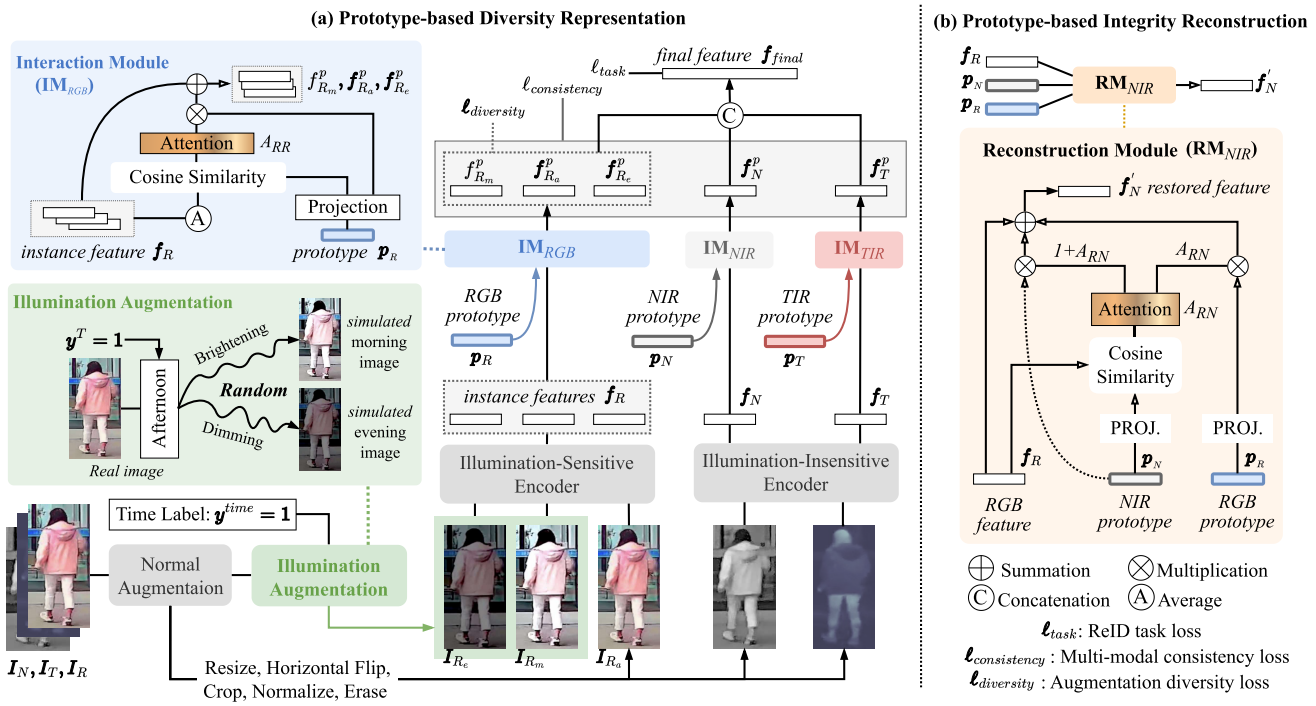


Fig. 2. The overview of the proposed Prototype-based Diversity and Integrity learning network (PDINet). (a) The RGB images will be augmented to diverse lighting conditions by illumination augmentation and sent to the illumination-sensitive encoder (independent of the NIR/TIR encoder). Then, the instance features interact with modality-specific prototypes to obtain the interactive features independently via interaction modules. (b) Taking the missing NIR modality as an example, we learn identity information from the existing RGB features and recover modality-specific information from the NIR prototype. By integrating their complementary information, we reconstruct the missing NIR modality.

interference and proposing adaptive recognition methods for dynamic light environments. However, existing pedestrian datasets still exhibit notable shortcomings in terms of data scale and scene complexity. Moreover, the unique challenges of vehicle ReID differ fundamentally from those of pedestrian ReID tasks. These limitations in current datasets restrict the generalization capability of existing pedestrian ReID algorithms in real-world complex scenarios.

III. Method

In this section, we will illustrate the proposed Prototype-based Diversity and Integrity learning network (PDINet). First, we will briefly outline the overall process and symbol definition in Sec. III-A. Then, we elaborate on the diversity representation phase, which includes the diverse augmentation of illumination-sensitive modality images and the interactive enhancement between multi-modal instance features and modal-specific prototypes in Sec. III-B. Finally, we elaborate on the integrity reconstruction phase, which leverages instance features containing identity information and prototype features containing modality information to recover the missing information in Sec. III-D.

A. Problem Definition

All-day multi-modal person re-identification aims to match the multi-modal images of the same person captured at different times. The overall structure of the proposed PDINet is illustrated in Fig. 2. In the training phase of our method, the input data is an image list $[I_{R_m}, I_{R_a}, I_{R_e}, I_N, I_T]$ with a time label y^{time}, camera label y^{cam} and identity label y^{id}. Note that $[I_{R_m}, I_{R_a}, I_{R_e}]$ are the result of the original RGB image I_R augmented by illumination augmentation. Then, lighting-sensitive (augmented RGB) and lighting-insensitive (NIR and TIR) images will be sent to the corresponding encoder to extract instance features $[f_R, f_N, f_T]$. Then, the instance features will interact with the modal-specific prototypes $[p_R, p_N, p_T]$ in three interaction modules designed for prototype-instance interaction. (1) In our framework, each modality is associated with a learnable prototype vector $p_c \in \mathbb{R}^{1\times D}$ where $D = 768$ corresponds to the dimensionality of the visual features. The set of all class prototypes is denoted as $P = [p_1; p_2; \ldots; p_C] \in \mathbb{R}^{C\times D}$. The prototype vectors p are initialized jointly by `nn.Parameter(torch.randn(C, D))` in PyTorch, ensuring sufficient diversity at the start of training. (2) During training, p is included in the optimization alongside all other network parameters: the loss function generates gradients for p, and all parameters (including p) are updated synchronously via backpropagation using the Adam optimizer. Finally, all features will be concatenated to form the final person descriptor f_{final}. In the aspect of loss constraints, the interacted features $[f^p_{R_m}, f^p_{R_a}, f^p_{R_e}, f^p_N, f^p_T]$ containing both modal and instance information will be constrained in subspace under the supervision of augmentation diversity loss $\ell_{diversity}$, multi-modal consistency losses $\ell_{consistency}$, and re-identification task loss ℓ_{task}.

B. Prototype-Based Diversity Representation

1) Illumination Augmentation: It is worth noting that NIR/TIR images benefit from special imaging principles and are robust to illumination changes, and only RGB images

may be influenced by the environment's lighting. Therefore, we only augment the RGB image to simulate the different lighting conditions at diverse periods. Following TransReID [32], we first implement the basic augmentations on each original input image I_R, including resize, random horizontal flip, zero padding, normalization, and random erasing. Then, to simulate the diverse illuminations, we randomly change the brightness of the RGB image after basic transformation. Since the original RGB image may come from different periods, we set four additional augmentation functions to brighten and dim the brightness. At the same time, to improve the diversity of lighting simulation, we set the upper and lower bounds (α, β) of the augmentation range and select a random value for each image. The selection of α and β is elaborated in Sec. VI-B. After setting up all the functions, illumination augmentation can easily convert the current image to different brightnesses according to the image's capture time label y^{time}. The final output of the augmentation is an image list $[I_{R_m}, I_{R_a}, I_{R_e}]$ with two simulated lighting images and one normal augmentation image. While illumination-based data augmentation has been commonly applied in person ReID, our proposed method differs from existing approaches in several key aspects. (1) Unlike common transform-based augmentations (e.g., brightness adjustment, which is typically used alongside random erasing, cropping, or padding) that apply random modifications to increase general data diversity, our method performs targeted illumination changes by simulating three specific time periods, aiming to improve temporal generalization. (2) Compared to low-light enhancement methods [33], [34] that often require additional training data [35], [36], [37], [38] and may suffer from dataset bias, our approach is lightweight and relies solely on the input image without external dependencies, thus offering better transferability and adaptability. (3) While some existing ReID approaches [39], [40] employ style augmentations (including brightness augmentation) at modality- or dataset-level to mitigate domain gaps. Since the time of query images is not fixed, our method instead focuses on improving robustness to temporal domain uncertainty rather than reducing explicit domain discrepancies.

Following the augmentation of the original RGB image, we will extract the multi-modal instance features to obtain their identity information. As we all know, only the RGB modality can be affected by lighting changes, while NIR and TIR are almost not influenced by ambient illumination, thanks to their imaging principles. Based on this observation, two independent encoders with the same structure are utilized to extract the lighting-sensitive (augmented RGBs) and the lighting-insensitive (NIR&TIR) features. We choose the ViT-base [19] pre-trained on ImageNet [41] as the feature encoder. Five instance features $[f_{R_m}, f_{R_a}, f_{R_e}, f_N, f_T] \in \mathbb{R}^{768\times 1}$ will be obtained after the initial image list $[I_{R_m}, I_{R_a}, I_{R_e}, I_N, I_T] \in \mathbb{R}^{3\times 256\times 128}$ extracted by lighting-sensitive encoder ϕ_{ls} and lighting-sensitive encoder ϕ_{li}, respectively. This processing can be formulated as:

$$\begin{aligned} f_i &= \phi_{ls}(I_i), i \in [R_m, R_a, R_e], \\ f_j &= \phi_{li}(I_j), j \in [N, T]. \end{aligned} \tag{1}$$

Moreover, we calculate the RGB center feature f_R by averaging the three RGB instance features, for subsequent interactions:

$$f_R = avg(f_{R_m}, f_{R_a}, f_{R_e}). \tag{2}$$

2) Interaction Module: Most previous multi-modal ReID methods only concentrate on the complementary information among modalities in the samples with the same identity, while ignoring modality-specific information at the dataset level. To obtain multi-modal domain information independent of identity, we initialized respective domain prototypes $p_R, p_N, p_T \in \mathbb{R}^{768}$ for RGB, NIR, and TIR. For later interaction and collaborative optimization, these domain prototypes are independent learnable vectors of the same size as the instance features. Previously extracted instance features will interact with modality prototypes to absorb the domain knowledge in the interaction module.

To enable instance features to take in domain information from modalities, we utilize the interaction between instance features and modality prototypes via the proposed interaction module (IM). For RGB modality, as shown in the left-top of Fig. 2 (a) IM_{RGB}, we first calculate the attention value A_{RR} between the RGB center feature and the RGB modal prototype via cosine similarity, and then the modal information in the prototype is integrated into the instance features through a two-step operation of multiplication and addition. We can express the IM_{RGB} mathematically as:

$$\begin{aligned} A_{RR} &= \vartheta(f_R, proj(p_R)), \\ f_i^p &= f_i + proj(p_R) \cdot A_{RR}, i \in [R_m, R_a, R_e], \end{aligned} \tag{3}$$

where $\vartheta(\cdot)$ denotes the cosine similarity calculation. $proj(\cdot)$ is a projection function consisting of several linear, batch normalization, and ReLU layers. For NIR and TIR modalities, the overall process of the interaction modules IM_{NIR} and IM_{TIR} is similar to IM_{RGB}, which also performs similarity calculation first and then performs weighted interaction. Mathematically, it can be represented as:

$$\begin{aligned} A_{NN} &= cos(f_N, proj(p_N)), \\ f_N^p &= f_N + proj(p_N) \times A_{NN}, \end{aligned} \tag{4}$$

$$\begin{aligned} A_{TT} &= cos(f_T, proj(p_T)), \\ f_T^p &= f_T + proj(p_T) \times A_{TT}. \end{aligned} \tag{5}$$

Different from the interaction strategies [42], [43], [44] in Incomplete Multi-view Clustering (IMC). As unsupervised clustering methods, they focus on discovering similarities among samples from the same class and mainly emphasize modality consistency, aiming to construct better multi-modal representations for shared subspace alignment. In contrast, our method targets the supervised multi-modal retrieval, where the emphasis lies in learning discriminative features. Furthermore, missing data is commonly handled using diffusion models or graph-based approaches [45], [46], [47], [48] in the field of IMC. In contrast, we leverage prototype-instance feature interaction to recover discriminative identity representations, which effectively compensates for the missing modality in the Prototype-based Integrity Reconstruction.

TABLE III
EXPERIMENTAL RESULTS OF BACKBONES UNDER DIFFERENT MODALITY COMBINATIONS ON THE ALLDAY843 AND ALLDAY843-G

Different Backbones		ResNet [61]		MobileNetV2 [62]		DenseNet [63]		ShuffleNet [64]		ViT [19]	
Evaluation Protocols		mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
Visible (RGB)		14.7	18.8	11.9	15.1	11.7	14.9	10.0	10.5	18.7	21.6
Near-infrared (NIR)		9.6	12.1	9.4	11.2	7.9	10.1	7.2	7.3	13.1	14.9
Thermal infrared (TIR)		6.4	6.1	6.7	8.1	7.3	9.4	5.1	6.2	9.7	11.6
AllDay843	*RGB+NIR*	23.8	32.8	16.0	17.5	20.2	26.8	14.1	17.2	35.1	40.0
	RGB+TIR	16.8	19.9	11.5	9.6	13.9	15.8	11.6	11.7	26.6	28.3
	RGB+NIR+TIR	**29.0**	**33.5**	**23.9**	**25.4**	**26.1**	**31.2**	**22.2**	**23.1**	**43.4**	**45.2**
AllDay843-G	*RGB+NIR*	20.7	29.9	14.9	18.0	18.0	24.1	12.0	15.2	31.2	36.5
	RGB+TIR	17.0	20.1	14.0	16.4	16.1	19.7	12.3	13.6	26.3	29.2
	RGB+NIR+TIR	**24.8**	**31.2**	**20.0**	**23.8**	**21.9**	**29.7**	**19.3**	**22.5**	**33.8**	**39.6**

After implementing the interaction between the instance feature and the corresponding modality prototype, both identity-independent modal information and identity-related instance information are present in the features $[f^p_{R_m}, f^p_{R_a}, f^p_{R_e}, f^p_N, f^p_T]$.

C. Prototype-Based Integrity Reconstruction

For the ReID task, where the training set and the test set have different identities, the pre-trained modality prototypes are essential because they have acquired the modality-specific knowledge that is unrelated to identity. Additionally, the complicated incomplete modality problem in all-day multi-modal ReID can be resolved with the aid of the domain characters provided by prototypes. In practical applications, the RGB modality is the easiest to obtain and provides the most information. Specifically, compared to NIR or TIR modalities, RGB images can provide rich color and texture information, which is crucial for recovering missing data. And the model achieves the highest accuracy when using the RGB modality alone (Table III). Therefore, when either NIR or TIR modality is missing, or both are absent, we utilize RGB features for reconstruction similar to the approach in [10]. Take missing NIR for example, we introduce the integrity reconstruction in Fig. 2 (b). Overall, the restored NIR feature is generated by the reconstruction module by sending the existing RGB instance feature f_R, the NIR prototype p_N, and the RGB prototype p_R. Specifically, we first calculate the attention A_{RN} of the current RGB instance and the NIR prototype, which aids in determining the relationship between the current RGB instance and the NIR domain. Then we restore part of the NIR instance information by multiplying the attention A_{RN} and the NIR prototype. Additionally, we will insert some RGB domain information into the NIR instance features to be restored since RGB modality information always exists. We can describe this process mathematically as:

$$\begin{aligned} A_{RN} &= 0.5 \times (1 + \vartheta(f_R, proj(p_N))), \\ f'_N &= f_R + (1 + A_{RN}) \times p_N + A_{RN} \times p_R. \end{aligned} \tag{6}$$

It is worth noting that even though the NIR prototype is pre-trained and fixed, each sample has unique RGB instance features, which means that the attention A_{RN} is unique, and the recovered information is distinct. The procedure for handling missing TIR is the same as that for missing NIR, and the formula is as follows:

$$\begin{aligned} A_{RT} &= 0.5 \times (1 + cos(f_R, proj(p_T))), \\ f'_T &= f_R + (1 + A_{RT}) \times p_T + A_{RT} \times p_R. \end{aligned} \tag{7}$$

The designed integrity reconstruction can rebuild distinct missing instance data for each incomplete sample using pre-trained prototypes and existing instance features. Furthermore, integrity reconstruction can be applied straight into the testing phase without training.

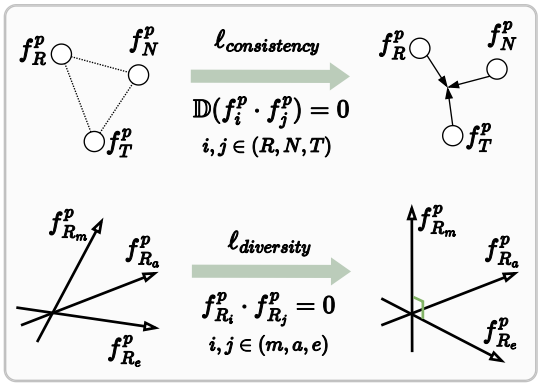


Fig. 3. Diagram of subspace constraint loss, including multi-modal consistency loss $\ell_{consistency}$ (top) and augmentation diversity loss $\ell_{diversity}$ (bottom).

D. Loss Functions

To train the proposed PDINet better, we use subspace constraint loss $\ell_{feature}$ (augmentation diversity loss $\ell_{diversity}$ and multi-modal consistency loss $\ell_{consistency}$) and ReID task loss ℓ_{task} (classification loss ℓ_{cls} and metric loss ℓ_{metric}). (1) Feature loss $\ell_{feature}$, as shown in Fig. 3. To guarantee the richness of the augmented information, there should be as much variation as possible between the augmented RGB features. Unit vectors have low similarity when they are orthogonal; to this end, we constrain all RGB features to remain augmentation diversity within the subspace through $\ell_{diversity}$. This goal can be expressed mathematically as:

$$\vec{f^p_{R_m}} \cdot \vec{f^p_{R_a}} = \mathbf{0}, \quad \vec{f^p_{R_a}} \cdot \vec{f^p_{R_e}} = \mathbf{0}, \quad \vec{f^p_{R_m}} \cdot \vec{f^p_{R_e}} = \mathbf{0}, \tag{8}$$

where $\cdot$ denotes the inner product of two vectors. The unit orthogonal vector can be expressed in space as two vectors

with an angle of 90 degrees. Therefore, $\ell_{diversity}$ can be defined as:

$$\ell_{diversity} = \vartheta(f^p_{R_m}, f^p_{R_a}) + \vartheta(f^p_{R_a}, f^p_{R_e}) + \vartheta(f^p_{R_m}, f^p_{R_e}), \tag{9}$$

where ϑ is the cosine similarity of two unit vectors:

$$\vartheta(f_1, f_2) = \frac{f_1 \cdot f_2}{max(\| f_1 \| \cdot \| f_2 \|, \epsilon)}, \tag{10}$$

where $\| \cdot \|$ denotes the norm of a vector, ϵ = 1e-8 is the default value to avoid division by zero.

To guarantee that the RGB modality has the same illumination stability as NIR or TIR, the multi-modal features should be consistent. Since unit vectors are highly similar when they are close in the subspace, multi-modal features are constrained to remain multi-modal consistent under the supervision of $\ell_{consistency}$. This goal can be represented mathematically as:

$$\begin{aligned} &\mathbb{D}(f^p_{R_m}, f^p_{R_a}) = 0, \\ &\mathbb{D}(f^p_{R_a}, f^p_{R_e}) = 0, \\ &\mathbb{D}(f^p_{R_m}, f^p_{R_a}) = 0, \end{aligned} \tag{11}$$

where $\mathbb{D}$ denotes the distance of two vectors. And we revise the hetero-center loss [49] in cross-modal ReID to $\ell_{consistency}$ for multi-modal ReID task as:

$$\begin{aligned} &f^p_R = avg(f^p_{R_m}, f^p_{R_a}, f^p_{R_e}) \\ &\ell_{consistency} = l_2(f^p_R, f^p_N) + l_2(f^p_R, f^p_T) + l_2(f^p_N, f^p_T), \end{aligned} \tag{12}$$

where l_2 denotes the mean squared error and can be formulated as:

$$l_2(f_1, f_2) = (f_1 - f_2)^2. \tag{13}$$

(2) ReID task loss ℓ_{task}. The widely used cross-entropy loss ℓ_{CE} is always employed in ReID to train a robust classifier and can be formulated as follows:

$$\ell_{CE} : (\tilde{y}, y) = -\sum_{i=1}^{B} y_i \log(\tilde{y}_i), \tag{14}$$

where B is the batch size, $\tilde{y}$ is the predicted results, y is the ground truth. In our training phase, we define two classification tasks: identity prediction and scene prediction. Specifically, the classification loss ℓ_{cls} can be formulated as:

$$\begin{aligned} &\ell_{cls} = \ell_{CE} : (cls(f_{final}), y^{id}) + \frac{1}{5}\sum_i \ell_{CE} : (cls(f^p_i), y^{cam}), \\ &where \quad i \in [R_m, R_a, R_e, N, T], \end{aligned} \tag{15}$$

where $cls(\cdot)$ represents the classifier layer, y^{id} and y^{cam} denotes the identity and camera label respectively. Triplet loss is also used in the training phase to constrain the feature distances, which can be formulated as follows:

$$\ell_{tri} : (f) =:: \sum_{i=1}^{P}\sum_{a=1}^{K}[m + \overbrace{\max_{p=1,\dots,K} D(f^i_a, f^i_p)}^{hardest:positive} - \overbrace{\min_{n=1,\dots,K} D(f^i_a, f^i_n)}^{hardest:negative}], \tag{16}$$

where f_a, f_p, and f_n denote the anchor, the positive, and the negative feature in the batch. In the training phase, we calculate the metric loss ℓ_{metric} as follows:

$$\begin{aligned} &\ell_{metric} = \ell_{tri} : (f_{final}) + \frac{1}{5}\sum_i \ell_{tri} : (f^p_i), \\ &where \quad i \in [R_m, R_a, R_e, N, T]. \end{aligned} \tag{17}$$

(3) Ultimately, we sum four losses as the final loss ℓ_{final}, which can be expressed mathematically as:

$$\begin{aligned} &\ell_{feature} = \ell_{diversity} + \ell_{consistency}, \\ &\ell_{task} = \ell_{cls} + \ell_{metric}, \\ &\ell_{final} = \omega_1 \times \ell_{feature} + \omega_2 \times \ell_{task}, \end{aligned} \tag{18}$$

where ω denotes the balance parameter for losses $\ell_{diversity}$ and $\ell_{consistency}$. The value of ω_1 used to balance subspace constraint loss $\ell_{feature}$ has been experimentally validated and set to 0.5. The weight of ℓ_{task} is set to ω_2 = 0.5 followed by the basic setting in TransReID [32].

IV. AllDay843 Dataset

In this paper, we capture multi-modal data from the real world and build a new AllDay843 dataset. The dataset includes all-day sequences captured by multiple cameras in diverse lighting conditions, integrating the complementarity of various modalities. The AllDay843 will provide a challenging benchmark for evaluating ReID methods and facilitate the development of robust and generalizable algorithms for real-world scenarios. In the following sections, we will describe the AllDay843 dataset, including the data collection and processing, dataset description, and comparison with existing datasets.

A. Data Collection and Processing

1) Data Collection: We collect all the raw video data of three modalities in two stages. In the first stage, we use paired Hikvision surveillance cameras with a resolution of 700 × 580 at 15 fps to record visible and near-infrared videos. At the same time, FLIR T640 cameras with a resolution of 640 × 480 at 20 fps are used to record thermal infrared videos. All cameras are operated simultaneously to ensure similar viewing angles and environments. In the second stage, we replace the paired Hikvision surveillance cameras with paired 360 AI cameras with a resolution of 2560 × 1440 at 24 fps. The thermal infrared videos are recorded by DALI S256 cameras with the same resolution and frame rate as FLIR T640 cameras. To ensure the diversity of pedestrian data and environments, we selected three places with dense crowds on campus to record the original videos, with each place containing two viewpoints. The dataset has a large shooting time, spanning spring, summer, and winter, and includes various weather conditions, such as sunny, cloudy, rainy, foggy, and snowy. The shooting time is not fixed, and data from morning, afternoon, and night are all included to provide a comprehensive dataset for ReID research. When collecting training data for the AllDay843 dataset, we did not require each individual to have data from multiple time

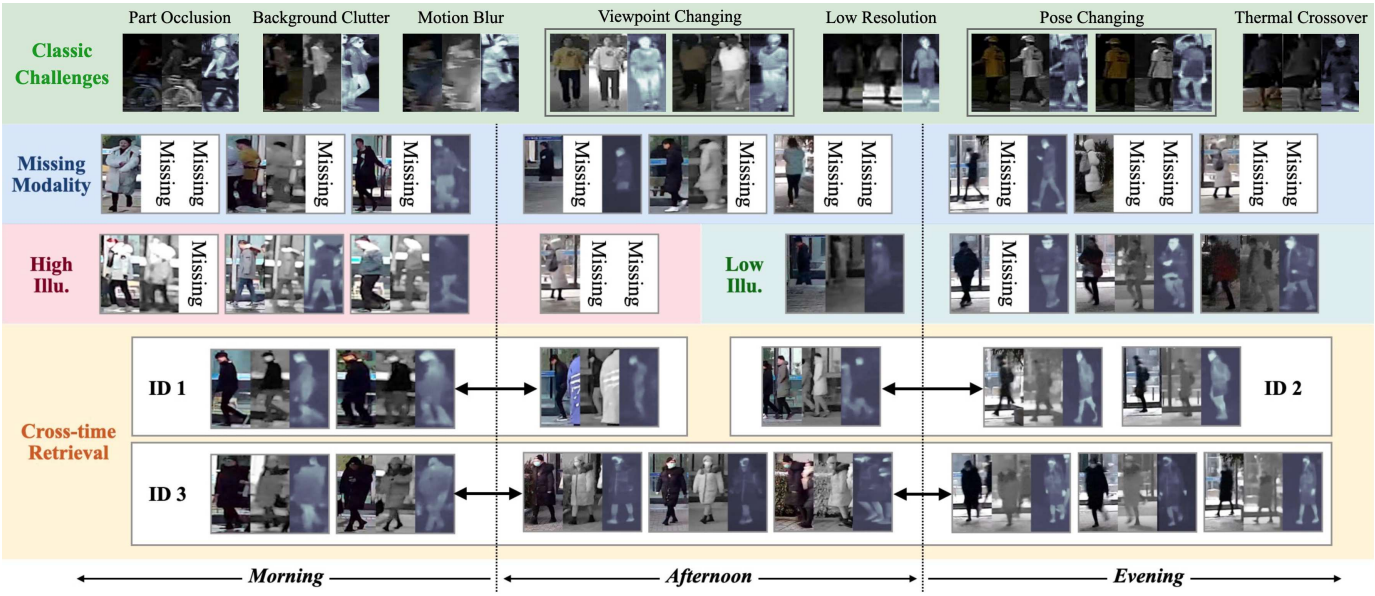


Fig. 4. Sample visualization of various challenges. Classic challenges in re-identification tasks, along with unique challenges introduced by multi-period conditions, include missing modality, high illumination (high illu.), low illumination (low illu.), and cross-time retrieval.

periods, so the difficulty of data collection is comparable to that of traditional multi-modal datasets.

2) Data Processing: After collecting the raw video data, we process them to meet the requirements of the multi-modal ReID task. Firstly, we synchronize the data from the three modalities to ensure that the pedestrian poses are almost aligned. Next, we export the video frames containing pedestrians and set a particular interval between the two exported frames to ensure pose and environment changes are captured. For images with complex backgrounds, the pedestrian image is manually cropped using the bounding box. For images with simple backgrounds, we use *yolo-v5* [50] to crop the pedestrian area. All samples in the final AllDay843 dataset are manually labeled with identity, shooting time, and view information. The appropriate information can be extracted from the image name for training. This rigorous data processing approach ensures that the AllDay843 dataset is of high quality and suitable for training and evaluating person re-identification algorithms.

B. Dataset Description

With the data collection and processing complete, we now provide a detailed description of the AllDay843 dataset, which contains 91,371 images of 843 identities captured at different times from 6 viewpoints with diverse illumination, weather, and background changes. We select 785 identities for training and 58 identities for testing. Most people have at least 100 images in each modality. The training set includes pedestrian images from the morning, afternoon, and night. Each identity in the test set contains images from two or three periods for all-day identification. To ensure a challenging evaluation, we introduce the all-day evaluation protocol during the testing phase, where each matched query and gallery image must be captured at different times and by different cameras. Moreover, AllDay843 includes a large number of image triplets for more comprehensive training and challenging testing, some samples of typical challenges are shown in Fig 4.

In addition, we construct two derived datasets, AllDay843-G and AllDay843-C, based on the AllDay843 dataset. (1) AllDay843-G Dataset. Existing ReID datasets do not take into account the issue of missing modalities, and current ReID methods are not applicable to multi-modal data with missing modalities. To enable the comparison with existing single-modality and multi-modal methods, we use the GAN-based Image Generation [51] to complete the missing data and extend AllDay843 into AllDay843-G. Compared to AllDay843, AllDay843-G maintains the same number of identities and cameras, while each sample has aligned three-modal images. (2) AllDay843-C Dataset. To compare with state-of-the-art cross-modal ReID methods, we reconstruct AllDay843-G into a cross-modal dataset called AllDay843-C. It includes three cross-modal settings: visible to infrared (AllDay843-C RN), visible to thermal (AllDay843-C RT) and infrared to thermal (AllDay843-C NT). These cross-modal test sets are constructed by removing the corresponding modalities from the query and gallery. Detailed settings of these datasets can be found in Table I.

C. Compared With Existing Datasets

Compared with existing prevalent ReID datasets, as shown in Table II, AllDay843 has the following major advantages:

1) AllDay843 Includes More Person Images Aligned in Three Modalities and Recorded by Six Non-Overlapping Cameras: The inclusion of multi-modal data in AllDay843 provides more complementary information than single-modality datasets. Additionally, the dataset contains more identities for training and testing than cross-modal datasets. When compared to other multi-modal person ReID datasets, AllDay843 has a significant advantage in terms of the number of identities

and images, making it an ideal benchmark for evaluating and improving state-of-the-art person ReID algorithms.

2) AllDay843 Includes Pedestrian Images From Three Time Periods and Five Weather Situations: The dataset contains pedestrian images captured in several time periods throughout the day, including in the morning, afternoon, and evening. It also includes images captured in different weather conditions, such as sunny, cloudy, rainy, foggy, and snowy, introducing many challenges to person ReID research. In addition to the usual challenges found in previous datasets, AllDay843 introduces new challenges caused by time shifts, such as clothing changes, diverse illumination, and thermal crossover. These factors make the dataset more representative of real-world scenarios.

3) AllDay843 Involves the Challenges Introduced by Multiple Modalities and Modality-Missing: It provides complementary information to traditional ReID tasks and offers new challenges for all-day multi-modal ReID research. The incorporation of multiple modalities in the dataset, even in the modality-missing condition, creates new opportunities to explore how different modalities can be effectively combined to improve person ReID performance, making AllDay843 a valuable resource for the person ReID research community.

V. Experiments

A. Experimental Data and Evaluation Metrics

1) Dataset: The dataset in the experimental section of this paper includes the proposed dataset AllDay843, as well as the derived datasets AllDay843-G and AllDay843-C.

2) Evaluation Metrics: We use the final features f_{final}, which are concatenated by all interacted features, as the person descriptor for similarity evaluation in the testing phase. The commonly used mean Average Precision (mAP) and Cumulative Matching Characteristic curve (CMC) are employed in our experiments to compare the performance of the proposed PDINet with other ReID methods. Additionally, due to the cross-time matching constraint in AMReID, we will only keep the samples that have the same identity with different cameras and time labels to form the ranking list. We use the Euclidean distance to measure the similarity of query and gallery features. Stated differently, the final samples in the ranking list meet three requirements set: distinct camera labels, the same identity labels, and different time labels.

B. Implementation Details

1) Basic Settings: PyTorch, with GeForce RTX 3090 GPUs, is the implementation platform. We employ the two vision transformers [19] as the encoders, these encoders are parameter independent and are pre-trained on Imagenet [41] to achieve higher performance with fewer iterations. In the basic augmentation (random crop, random erase, and random horizontal flip) and proposed illumination augmentation, images are resized to unify size (256×128). The mini-batch size is set to 32, and the maximum epoch is set to 60. More details can be found in our code.

2) Optimization: The widely used Stochastic Gradient Descent (SGD) with a momentum of 0.9 and weight decay of 0.0001 is employed to fine-tune our model. The initial learning rate was established at 0.001. The dim of modality prototypes is 768-dim, which is the same as the dimension of instance features extracted by the encoders. The final feature scale for every individual is $768 \times 5 = 3840$-dim. The balance parameter ω for $\ell_{diversity}$ and $\ell_{consistency}$ is set as 0.5 in the final loss. The lower and upper bounds of the augmentation range α and β are set as 0.1 and 0.4, respectively.

C. Necessity of Multi-Modal Data

In this section, we investigate the importance and necessity of using multi-modal data for all-day person ReID. Accordingly, we compare the results obtained by various base feature extraction networks for different modality combinations, as presented in Table III. We select four CNN-based backbones [61], [62], [63], [64] and the transformer-based framework, ViT [19]. Since the AllDay843 dataset includes samples with missing modalities, we integrate zero-padding to replace the modality-missing images for all comparison methods.

1) Results of Single-Modality: Due to its powerful extraction ability on global features, ViT outperforms traditional CNN-based frameworks and achieves the highest accuracy.

2) Results of Dual-Modality: The results on AllDay843 show that the incorporation of additional NIR information can lead to further enhancements in the accuracy of RGB single-modality. For RGB+TIR, the TIR modality can improve both mAP and Rank-1 simultaneously. The improvement in accuracy in both cases comes from the texture details and temperature information complementary to RGB in the NIR and TIR modalities.

3) Results of Multi-Modal: Moreover, incorporating both NIR and TIR to support the available RGB can lead to the highest accuracy, which varies with different backbone architectures. The mAP can be improved by at least 12.0% (achieved by MobileNetV2 [62]) and up to 24.7% (achieved by ViT [19]). This superior result shows that the information of NIR and TIR not only complements RGB but also is different from each other.

Similar accuracy improvements can also be observed in the AllDay843-G dataset. The AllDay843-G dataset comprises complete multi-modal data, so we only need to extract features from different modalities and concatenate them to obtain the final representation. Whether RGB is combined with NIR or TIR, it can achieve higher mAP and Rank-1 accuracy than RGB alone. Furthermore, when using the RGB+NIR+TIR, each backbone can achieve the highest mAP and Rank-1.

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

As shown in Table IV, we compare the proposed PDINet with state-of-the-art single- and multi-modal ReID methods on AllDay843 and AllDay843-G, including CNN-based [3], [10], [15], [65], [66], [67], [68] and ViT-based [13], [14], [16], [17], [18], [21], [32]. IDEA [21] employs additional text, we report the results using the official model (without the text) for a fair comparison with other methods. First, the proposed PDINet

TABLE IV

EXPERIMENTAL RESULTS COMPARING PDINET WITH STATE-OF-THE-ART METHODS ON ALLDAY843 AND ALLDAY843-G. * DENOTES THE METHOD IS SPECIALLY DESIGNED FOR MULTI-MODAL REID

Testing Dataset / Evaluation Protocols	Publication	Backbone	*AllDay843* mAP	*AllDay843* Rank-1	*AllDay843* Rank-5	*AllDay843* Rank-10	*AllDay843-G* mAP	*AllDay843-G* Rank-1	*AllDay843-G* Rank-5	*AllDay843-G* Rank-10
PCB [65]	ECCV2018	CNN	26.7	28.0	38.4	44.8	22.7	26.0	36.6	42.3
MLFN [66]	CVPR2018	CNN	24.0	23.5	36.9	45.9	18.2	20.5	32.9	40.4
OSNet [67]	ICCV2019	CNN	22.7	24.5	35.0	42.3	17.7	23.0	33.7	39.5
StrongBaseline [68]	CVPRW2019	CNN	27.6	34.2	44.7	51.3	24.0	32.4	44.8	52.3
AGW [3]	TPAMI2021	CNN	28.0	32.4	43.6	49.4	26.6	33.5	44.5	50.8
PFNet* [10]	AAAI2021	CNN	25.1	29.5	38.6	44.3	21.5	26.2	37.0	43.9
IEEE* [15]	AAAI2022	CNN	23.2	26.7	31.8	34.9	20.3	26.5	33.8	38.0
TransReID* [32]	ICCV2021	ViT	37.0	43.4	52.0	**56.6**	**33.6**	**39.6**	**49.2**	**54.2**
HTT* [16]	AAAI2024	ViT	**40.6**	**44.6**	**54.2**	**59.6**	32.0	37.4	**47.3**	**52.6**
TOP-ReID* [13]	AAAI2024	ViT	27.0	34.3	40.2	43.7	**35.1**	**39.0**	46.0	50.1
EDITOR* [17]	CVPR2024	ViT	34.3	41.4	51.4	55.8	27.3	33.3	42.8	47.6
DEMO* [18]	AAAI2025	ViT	**41.2**	**44.3**	**56.9**	**59.6**	29.5	33.1	43.0	47.9
MambaPro* [14]	AAAI2025	ViT	33.0	33.6	42.8	47.1	27.9	30.8	39.6	44.2
IDEA* [21]	CVPR2025	ViT	27.6	29.2	35.4	39.3	27.9	30.6	40.3	45.0
PDINet*	Ours	ViT	**44.8**	**49.0**	**60.3**	**65.6**	**42.4**	**47.7**	**59.3**	**64.7**

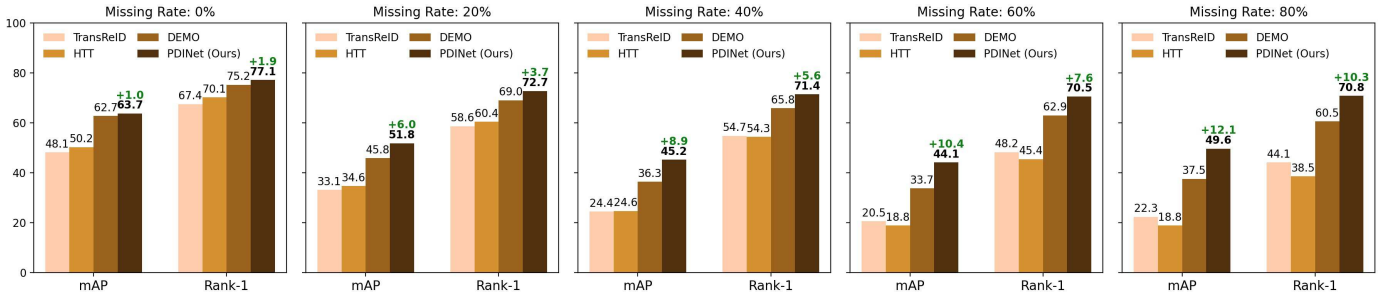


Fig. 5. Comparison with TransReID [32], HTT [16] and DEMO [18] on Market1501-AllDay with varying random missing rates.

outperforms existing methods across all metrics (mAP, Rank-1, Rank-5, Rank-10) on both AllDay843 and AllDay843-G datasets, especially among models using ViT as the backbone. Second, models with ViT as the backbone generally perform better than CNN-based models in terms of Rank-1 and Rank-5 on AllDay843 and AllDay843-G. Moreover, performance drops across all methods in AllDay843-G; this indicates that AllDay843-G is more challenging, likely due to the significant noise in the restored images, which makes feature extraction and matching more difficult. In contrast, PDINet shows minimal accuracy degradation. This can be attributed to the constraints on different modality features and stable prototype-based information interaction, which helps ensure more robust feature alignment and mitigates the impact of noise in restored images. As a result, PDINet achieves strong generalization across different conditions, maintaining high accuracy even in more challenging scenarios.

Additionally, since existing multi-modal person ReID dataset [10] fails to provide explicit time labels. Consequently, to rigorously evaluate the generalization capability on cross-time retrieval of our proposed method, we reconstruct the simulated all-day dataset, Market1501-AllDay. First, we simulate images captured in the afternoon and evening by reducing brightness and altering the colors of RGB images, while assigning corresponding temporal labels. Second, for the NIR and TIR modalities, we retain the data from Market1501-MM [15]. The final Market1501-AllDay thereby meets cross-time retrieval requirements. We conduct experiments comparing the proposed PDINet with the top three multi-modal methods (TransReID [32], HTT [16], and DEMO [18]) on Market1501-AllDay, and also test the performance under different missing ratios. As shown in Fig. 5, all methods achieve generally higher accuracy on Market1501-AllDay than on AllDay843. This is primarily because the time-period variations in Market1501-AllDay are synthetically generated, making it difficult to match the diverse environmental factors present in real-world scenarios, thus resulting in lower cross-time retrieval difficulty. Meanwhile, our method demonstrates consistent improvements across all experimental settings compared to other approaches (with mAP gains of 1.0%-12.1% and Rank-1 improvements of 1.9%-10.3% over the suboptimal DEMO [18]), highlighting its superior adaptability. Due to the design of integrity reconstruction, our method exhibits stronger performance advantages over DEMO [18], with the gap widening as the missing ratio increases. At an 80% missing ratio, the leads in mAP and Rank-1 reach 12.1% and 10.3%, respectively.

E. Comparison With Cross-Modal Methods

Different modalities exhibit varying image quality at different times of day, thus the retrieval performance may be improved by using only the modality that performs best under the current time condition. This idea aligns with the

TABLE V
EXPERIMENTAL RESULTS OF TAENET COMPARING WITH STATE-OF-THE-ART CROSS-MODAL METHODS ON ALLDAY843-C

Testing Mode	*Visible to Infrared*			*Infrared to Visible*			*Visible to Thermal*			*Thermal to Visible*			*Infrared to Thermal*			*Thermal to Infrared*		
Evaluation Protocols	mAP	R-1	R-5	mAP	R-1	R-5	mAP	R-1	R-5	mAP	R-1	R-5	mAP	R-1	R-5	mAP	R-1	R-5
DDAG [68]	9.8	11.8	21.9	10.1	12.5	20.5	5.3	7.8	14.3	**5.4**	5.8	12.7	4.0	4.4	10.0	4.0	4.5	8.9
AGW [3]	9.4	13.3	22.6	10.0	12.9	21.3	4.4	4.4	9.4	4.2	3.9	10.0	3.6	3.7	10.4	3.5	4.3	10.0
CAJ [40]	**11.4**	**14.6**	**25.1**	**11.5**	**15.1**	**25.8**	5.3	7.6	15.8	4.3	4.0	10.0	3.7	5.2	10.6	3.9	4.3	8.5
LbA [69]	8.8	11.7	20.2	9.3	12.5	22.5	4.0	4.7	10.0	4.4	5.3	11.3	**5.6**	**7.2**	13.8	**4.9**	**6.2**	**13.3**
PIC [70]	9.0	11.8	20.0	9.0	12.9	19.9	**5.5**	5.2	10.9	**5.8**	5.1	10.0	4.8	5.7	11.2	3.4	3.6	7.5
DEEN [5]	7.3	10.2	19.0	9.0	**13.6**	22.3	5.2	**8.6**	**18.8**	4.9	**6.0**	14.0	**5.5**	**7.7**	**14.0**	3.8	4.3	9.0
IDKL [71]	8.8	**12.9**	**29.4**	8.3	10.2	**29.3**	4.6	4.3	12.8	5.0	5.7	**19.3**	4.7	4.5	**16.1**	4.1	4.3	**14.5**
PDM [72]	10.0	11.0	20.5	**10.3**	12.7	22.5	**6.2**	**9.7**	**18.0**	5.2	**8.0**	**15.6**	4.9	4.9	9.0	**5.7**	**7.1**	11.7
HHRG [74]	**10.6**	**12.9**	19.6	10.0	11.9	20.1	5.2	6.6	11.5	4.5	4.0	7.5	5.2	6.4	11.6	**4.9**	5.4	10.7
PDINet (dual-modal)	mAP: **31.1**		R-1: **35.0**		R-5: **46.7**		mAP: **32.2**		R-1: **35.9**		R-5: **46.1**		mAP: **24.8**		R-1: **29.8**		R-5: **41.1**	

TABLE VI
ABLATION STUDY OF PROPOSED METHOD ON ALLDAY843-G AND ALLDAY843

Datasets		Diversity Representation			Integrity Reconstruction	Evaluation Metrics			
		Augmentation	Constraint	Interaction		mAP	R-1	R-5	R-10
AllDay843-G	*a*	-	-	-	-	34.9	39.6	50.5	56.2
	b	✓	-	-	-	39.0	44.4	55.8	61.3
	c	✓	✓	-	-	42.3	45.7	57.5	62.3
	d	✓	✓	✓	-	**42.4**	**47.7**	**59.3**	**64.7**
AllDay843	*e*	-	-	-	-	34.5	38.8	50.4	55.6
	f	✓	-	-	-	39.7	44.0	55.8	61.5
	g	✓	✓	-	-	41.7	46.1	57.7	63.3
	h	✓	✓	✓	-	42.2	47.4	59.0	64.5
	i	✓	✓	✓	✓	**44.8**	**49.0**	**60.3**	**65.6**

original goal of cross-modal ReID, which achieves robust feature matching across different modalities. To investigate the necessity of multi-modal ReID in comparison to the cross-modal scenario, we create the cross-modal setting. Specifically, we use visible-thermal images to replicate the modality setting of RegDB [55] and visible-infrared images to imitate the SYSU [2], and additionally introduce an infrared–thermal pairing to achieve a more comprehensive modality combination. To maintain a fair comparison, we evaluate PDINet on the corresponding dual-modality scenarios. We compare PDINet with nine state-of-the-art cross-modal person ReID methods [3], [5], [40], [69], [70], [71], [72], [73], [74], as presented in Table V.

First, cross-modal ReID (on AllDay843-C) experiences a significant performance drop compared to multi-modal ReID (on AllDay843 or AllDay843-G). Even weaker methods achieve mAP above 20% on the AllDay843 dataset, while ViT-based models reach 30-40%. In the cross-modal ReID, the mAP generally drops to 5%-15%. Second, in the visible-infrared task, CAJ [40] performs best in the visible-infrared tasks. Moreover, in the visible-thermal and infrared-thermal scenario, performance is even lower, with most Rank-1 scores below 10%, due to the larger inherent imaging differences between thermal and the other modalities.

In contrast, PDINet (dual-modality) utilizes the complementary information between different modalities of samples and also takes advantage of the stable information from the entire dataset's modality domain, thus, it achieves the highest accuracy on both visible-infrared and visible-thermal scenarios. Specifically, for visible-infrared, PDINet (mAP: 31.1%) far exceeds the highest value of CAJ [40] with an improvement margin of 19.6%. And PDINet (Rank-1: 35.0%) is about 2.31 times higher than the second-best method CAJ [40]. The experimental results show that existing cross-modal approaches failed to significantly improve performance under cross-time retrieval. First, AllDay843-C is not a typical cross-modal dataset, the dominant modality varies with the query time, whereas existing methods use a fixed extractor per modality and can't adapt. Second, different modalities provide complementary information under varying temporal conditions that current cross-modal approaches overlook, thereby limiting their performance. Overall, methods that rely on one heterogeneous modality struggle in real-world, cross-time retrieval. Our results underscore the need for joint multi-modal modeling.

F. Ablation Study

Extensive experiments are conducted on AllDay843 and AllDay843-G to evaluate the effectiveness of each component in PDINet, as shown in Table VI. From the results, the addition of each module can improve the values of various evaluation metrics. For both AllDay843 and AllDay843-G, when all modules are used together, the evaluation metrics reach relatively the highest. This indicates that the collaborative work of multiple modules has a positive effect and can effectively enhance performance on the datasets.

(1) **Effect of illumination augmentation.** To simulate the various lighting conditions, we propose the illumination augmentation to brighten or dim the image. From the results in lines *b* and *f*, we find that illumination augmentation is effective in solving cross-time matching challenges, and improved mAP by 4.1% and 5.2%, and Rank-1 by 4.8% and 5.2% on the

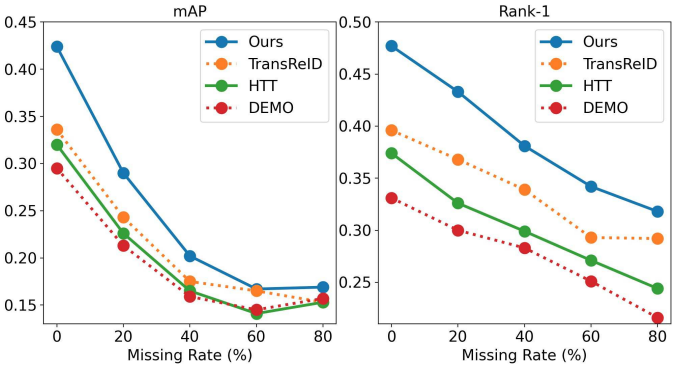


Fig. 6. Comparison with TransReID [32], HTT [16] and DEMO [18] under varying random missing rates.

two datasets, respectively. (2) **Effect of subspace constraint loss.** To ensure the modal consistency and scene richness in extracted features, we propose the subspace constraint loss to constrain them. Based on the data from lines *c* and *g*, we observe that the mAP is higher than using SLA alone, reaching 42.3% and 41.7% on both datasets, and Rank-1 is also improved by 1.3% and 2.1%. (3) **Effect of interaction.** To help the instance features absorb the modality-specific information, we define three modality prototypes and interact them with the corresponding instance. According to the data in lines *d* and *h*, it can be seen that mAP and Rank-1 continue to improve. (4) **Effect of integrity reconstruction.** The proposed integrity reconstruction can recover unique features based on the specifics of each missing instance. As can be seen from the results in line *i*, when testing with the complete PDINet, mAP, and Rank-1 reached their highest, which are 44.8% and 49.0%, respectively.

VI. Discussions

A. Discussion on Random Missing Rate

To simulate the diverse random missing scenarios encountered in real-world applications and to evaluate the robustness of our method under such conditions, we conduct additional experiments on the test set by discarding modalities at rates of 20%, 40%, 60%, and 80%. For each missing rate, we apply independent Bernoulli masks and compare our approach with the three highest-accuracy Transformer-based multi-modal methods [16], [18], [32], the results are shown in Fig 6. As the missing rate increases from 0% (without missing) to 80%, the mAP and Rank-1 accuracy of all methods steadily decline, indicating that modality loss degrades ReID performance; however, our model (blue solid line) consistently achieves the highest accuracy and the most gradual performance drop at every missing level, outperforming the three Transformer-based baselines (orange, green, and red curves) by 2.6% percentage points on Rank-1, thereby demonstrating superior robustness and stability across various random missing scenarios.

B. Discussion on Illumination Augmentation

1) Augmentation Range Selection: In the proposed illumination augmentation, we brighten or dim the image via the true time label. It is worth noting that we do not fix the degree of brightness or dimness but randomly choose a value within a certain range. Such random augmentation can improve the brightness richness of the augmented images and thereby improve the network's robustness to illumination. The choice of upper and lower bounds is based on the following two main considerations: (1) Lighting changes are typically influenced by the day-night cycle, weather conditions, and environmental factors. To simulate real-world lighting variations, we chose representative lighting characteristics for three periods (morning, afternoon, and evening) to correspond to morning, afternoon, and evening intensity ranges, respectively. (2) During augmentation for each specific time period, we randomly sample lighting intensity values to increase augmented data diversity. Overall, applying random values within these intervals both reflects real-world lighting patterns and balances data diversity with model robustness. We conduct experiments to find a suitable change range, as shown in Table VII.

TABLE VII
Comparison Results of Using Different Augmentation Ranges in Illumination Augmentation on AllDay843-G

Δ	*lower boundary*	*upper boundary*	mAP	Rank-1	Rank-5	Rank-10
0.2	0.1	0.3	41.9	45.7	57.3	62.2
	0.2	0.4	41.7	47.3	58.2	62.8
	0.3	0.5	40.4	46.8	58.0	63.0
0.3	0.1	0.4	**42.4**	47.7	**59.3**	**64.3**
	0.2	0.5	41.5	**48.4**	59.1	63.3
0.4	0.1	0.5	40.4	46.6	57.2	62.2

TABLE VIII
Comparison Results of Using Different Augmentation Methods on AllDay843-G

Low-light Image Enhancement Methods	Extra Training Datasets	AllDay843		AllDay843-G	
		mAP	Rank-1	mAP	Rank-1
Retinexformer [33]	LOL_v1 [35]	35.4	39.3	36.6	40.7
	LOL_v2_real [36]	34.5	38.0	36.0	41.2
	SID [37]	34.9	38.5	37.0	41.6
	SMID [38]	33.7	38.1	34.5	39.5
HVI-CIDNet [34]	LOL_v1 [35]	34.5	35.3	34.0	36.2
	LOL_v2_real [36]	34.1	35.8	34.1	36.9
	SID [37]	34.9	36.5	33.9	35.6
Two-period Simulation	\	41.2	44.3	40.5	44.4
Three-period Simulation	\	**44.8**	**49.0**	**42.4**	**47.7**

The range Δ is determined by the changing lower bound α and upper bound β. It can be seen from Table VII that the best effect is achieved when the $\Delta = 0.3$, and the highest values of each indicator are obtained at this time. This is mainly because if Δ is too small, the changes of brightness may not be diverse enough, and if Δ is too large, the stability of the training may be affected. After determining the range and taking comprehensive considerations into account, we used a combination of $\alpha = 0.1$ and $\beta = 0.4$ in the experiment.

2) Different Augmentation Methods: To validate the simplicity and effectiveness of the proposed illumination augmentation, we also compared it with low-light enhancement methods. It is important to note that such methods typically

TABLE IX
COMPARISON RESULTS OF THE EFFECTS OF DIFFERENT LOSSES IN FEATURE LOSS ℓ_{FC} ON ALLDAY843-G

Method	mAP	Rank-1	Rank-5	Rank-10
ℓ_{FC}	**42.4**	**47.7**	**59.3**	**64.7**
w/o $\ell_{diversity}$	36.3	40.2	50.5	56.4
w/o $\ell_{consistency}$	38.5	46.2	56.9	62.3

learn illumination transformation relationships from paired normal and low-light image sets, meaning they rely on additional datasets for training [35], [36], [37], [38]. Here, we choose two methods, Retinexformer [33] and HVI-CIDNet [34], and apply their pre-trained models (trained on multiple additional datasets) to enhance the low-light RGB images in AllDay843. The results are shown in Table VIII.

Across all metrics, the three-period simulation further improves performance compared to the two-period simulation, indicating that a more comprehensive temporal simulation can enhance the model's adaptability to all-day data variations. Overall, our strategy achieves better results than low-light enhancement methods without relying on additional data. The possible reason is that the distribution of the datasets used for training low-light enhancement methods may differ significantly from that of AllDay843-G, limiting the generalization ability of these models. Moreover, these low-light enhancement methods primarily focus on improving image brightness and visual quality rather than directly optimizing the retention and discrimination of features for person ReID tasks. In contrast, our strategy is directly optimized on the ReID task data, allowing for better preservation of critical information.

C. Discussion on Subspace Constraint Loss

To further verify the role of $\ell_{diversity}$ and $\ell_{consistency}$, we remove two losses respectively and observed the fluctuation of the results, as shown in Table IX. First, when two losses work together, the highest mAP and Rank-1 values can be obtained. Secondly, after removing the $\ell_{diversity}$ that constrains the augmented RGB features, the mAP drops by 6.1%, and the Rank-1 value drops by 7.5%, indicating that increasing the richness of the augmented features through $\ell_{diversity}$ is beneficial. Meanwhile, after removing the $\ell_{consistency}$ that constrains different modality features, mAP and Rank-1 decrease by 3.9% and 1.5%, respectively, proving that improving the similarity of multi-modal features is helpful to the discrimination of the final features.

In addition, the radar chart (Fig. 7) compares the results of using different distance metrics (l_1, l_2, *cos*) in $\ell_{diversity}$ loss. As can be seen from the chart, in the mAP metric, the *cos* distance metric is comparable to the l_2 distance metric, but both perform better than l_1, which has a value of 37.0%. In the Rank-n metric, the *cos* distance metric consistently outperforms both l_1 and l_2 metrics. Ultimately, we choose to use the *cos* distance metric in $\ell_{diversity}$ loss.

Moreover, the line chart (Fig. 8) shows the experimental results on the AllDay843-G dataset when the parameter ω in

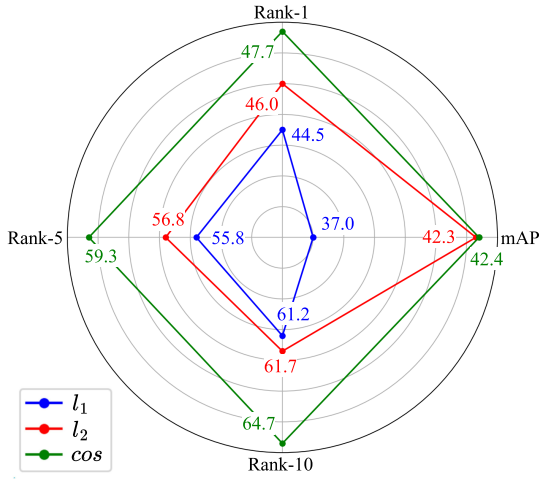


Fig. 7. Comparison results of using different distance metrics in $\ell_{diversity}$ on AllDay843-G.

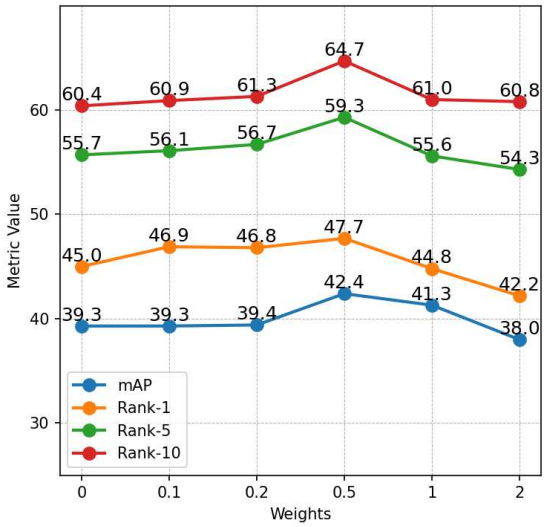


Fig. 8. Results of different values of parameter ω in the final loss ℓ_{final} on AllDay843-G.

the final loss function ℓ_{final} takes different values (from 0 to 2). It can be observed that for the Rank-n metric, the highest values are achieved when $\omega = 0.5$, after which it decreases. The mAP metric reaches its highest value of 42.4% when $\omega = 0.5$. These results indicate that the parameter ω affects the model's performance, with $\omega = 0.5$ achieving optimal performance across various metrics.

D. Discussion on Integrity Reconstruction

To verify the effectiveness of the integrity reconstruction compared with various information reconstruction methods, we conduct experiments in this section for discussion. (1) Padding reconstructions address the issue of missing images by employing randomly generated tensors (random padding) and zero tensors (zero padding [76]), ensuring that their dimensions match the other existing inputs. In PDINet, we generate a matrix matching the input dimensions to construct a "complete" multi-modal input used for training and testing.

TABLE X
COMPARISON RESULTS OF INTEGRITY RECONSTRUCTION AND OTHER METHODS FOR INCOMPLETE MODALITY ISSUE ON ALLDAY843

Method	mAP	Rank-1	Rank-5	Rank-10
Only RGB	18.7	21.6	30.1	35.2
Random Padding	41.4	44.7	55.3	60.3
Zero Padding	42.6	45.8	56.2	61.0
Image Rec.				
cycleGAN [51]	35.6	40.8	49.8	54.5
Feature Rec.				
Wang et al. [13]	40.0	43.9	55.3	60.3
Wang et al. [10]	44.1	46.4	56.8	61.8
Integrity Reconstruction	**44.8**	**49.0**	**60.3**	**65.6**

(2) Image reconstruction generate new pixels that resemble the real data to ensure the reconstructed output matches the real modality in both visual details and statistical characteristics. We use the same CycleGAN [51] architecture and loss weights as in [15], we train a generator on paired multi-modal data to produce the missing images, creating the AllDay843-G dataset, which is then fed into PDINet. (3) Feature reconstruction learn to map intermediate features into the feature space of the missing modality and reconstruct the missing information while preserving semantic consistency. Inspired by the classic CNN-based and ViT-based methods [10], [13], we integrate their feature reconstruction module into PDINet and train them jointly, then use the reconstructed features for ReID during testing. These two reconstruction modules are based on convolutional layers and multi-head attention designs, respectively.

First, the results of all methods designed for modality-missing, as shown in Table X, surpass the single-modality approach. However, the image reconstruction method fails to achieve satisfactory results compared to the other methods due to the lack of sufficient multi-modal training data for effective generator training. Second, although both random padding and zero padding yield comparable results, random padding introduces unpredictability, leading to unstable performance. In contrast, zero padding does not contribute useful information and cannot leverage the advantages of multi-modal data training. Moreover, feature reconstruction similar to [10] can reconstruct missing information while harnessing the benefits of multi-modal fusion. In contrast, the proposed incomplete information restoration achieves the highest accuracy among all the compared methods due to the recovered features that combine both the instance information of the sample itself and the stable modal prototype information, ensuring that the reconstructed information is unique and identity-discriminative.

VII. CONCLUSION

In this paper, we work on the challenge of all-day multi-modal person re-identification, which leverages multi-modal data for pedestrian retrieval across multiple periods. To tackle cross-time retrieval and modality-missing challenges, we propose the Prototype-based Diversity and Integrity learning network (PDINet). The diversity representation augments lighting-sensitive images and enriches all features via model-specific prototype interaction, while integrity reconstruction recovers missing modality data using prototype-guided relationships. Moreover, we create a comprehensive benchmark dataset for this task, called AllDay843, which consists of 91,371 images of 843 different identities, and involves real-world cross-time retrieval and modality missing challenges. Extensive experiments show that PDINet performs much better than the state-of-the-art methods on the novel AllDay843 dataset and its derived datasets. In the future, we will continue researching the quality of imbalanced multi-modal data, considering modal quality assessments to address the variability in data quality caused by different lighting conditions and environmental factors at various times. And we will consider more missing scenarios, such as how to leverage NIR/TIR modalities without color information to recover the texture and surface details of RGB images when the RGB modality is missing.

REFERENCES

[1] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1116–1124.

[2] 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. 5380–5389.

[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] H. Lu, X. Zou, and P. Zhang, "Learning progressive modality-shared transformers for effective visible-infrared person re-identification," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, Washington, DC, USA, 2023, pp. 1835–1843.

[5] 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)*, Vancouver, BC, Canada, Jun. 2023, pp. 2153–2162.

[6] F. Zhu, Y. Zhu, X. Jiang, and J. Ye, "Cross-domain attention and center loss for sketch re-identification," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 3421–3432, 2022.

[7] A. Lu, C. Li, T. Zha, X.-F. Wang, J. Tang, and B. Luo, "Nighttime person re-identification via collaborative enhancement network with multi-domain learning," *IEEE Trans. Inf. Forensics Security*, vol. 20, pp. 1305–1319, 2025.

[8] H. Nguyen, K. Nguyen, S. Sridharan, and C. Fookes, "AG-ReID.V2: Bridging aerial and ground views for person re-identification," *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 2896–2908, 2024.

[9] S. Chen, M. Ye, Y. Huang, and B. Du, "Towards effective rotation generalization in UAV object re-identification," *IEEE Trans. Inf. Forensics Security*, vol. 20, pp. 2593–2606, 2025.

[10] A. Zheng, Z. Wang, Z. Chen, C. Li, and J. Tang, "Robust multi-modality person re-identification," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2021, vol. 35, no. 4, pp. 3529–3537.

[11] H. Li, C. Li, X. Zhu, A. Zheng, and B. Luo, "Multi-spectral vehicle re-identification: A challenge," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2020, vol. 34, no. 7, pp. 11345–11353.

[12] A. Zheng, X. Zhu, Z. Ma, C. Li, J. Tang, and J. Ma, "Cross-directional consistency network with adaptive layer normalization for multi-spectral vehicle re-identification and a high-quality benchmark," *Inf. Fusion*, vol. 100, Dec. 2023, Art. no. 101901.

[13] Y. Wang, X. Liu, P. Zhang, H. Lu, Z. Tu, and H. Lu, "TOP-ReID: Multi-spectral object re-identification with token permutation," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2023, pp. 5758–5766.

[14] Y. Wang et al., "MambaPro: Multi-modal object re-identification with mamba aggregation and synergistic prompt," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 39, 2025, pp. 8150–8158.

[15] Z. Wang, C. Li, A. Zheng, R. He, and J. Tang, "Interact, embed, and enlarge: Boosting modality-specific representations for multi-modal person re-identification," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 36, no. 3, 2022, pp. 2633–2641.

[16] Z. Wang, H. Huang, A. Zheng, and R. He, "Heterogeneous test-time training for multi-modal person re-identification," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 38, 2024, pp. 5850–5858.

[17] P. Zhang, Y. Wang, Y. Liu, Z. Tu, and H. Lu, "Magic tokens: Select diverse tokens for multi-modal object re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 17117–17126.

[18] Y. Wang, Y. Liu, A. Zheng, and P. Zhang, "Decoupled feature-based mixture of experts for multi-modal object re-identification," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2025, pp. 8141–8149.

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

[20] A. Zheng, Z. Ma, Y. Sun, Z. Wang, C. Li, and J. Tang, "Flare-aware cross-modal enhancement network for multi-spectral vehicle re-identification," *Inf. Fusion*, vol. 116, Apr. 2025, Art. no. 102800.

[21] Y. Wang, Y. Lv, P. Zhang, and H. Lu, "IDEA: Inverted text with cooperative deformable aggregation for multi-modal object re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2025, pp. 29701–29710.

[22] C. Li, X. Liang, Y. Lu, N. Zhao, and J. Tang, "RGB-T object tracking: Benchmark and baseline," *Pattern Recognit.*, vol. 96, Dec. 2019, Art. no. 106977.

[23] L. Zhang, A. Gonzalez-Garcia, J. van de Weijer, M. Danelljan, and F. S. Khan, "Synthetic data generation for end-to-end thermal infrared tracking," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1837–1850, Apr. 2019.

[24] T. Zhang, H. Guo, Q. Jiao, Q. Zhang, and J. Han, "Efficient RGB-T tracking via cross-modality distillation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 5404–5413.

[25] T. Hui et al., "Bridging search region interaction with template for RGB-T tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 13630–13639.

[26] T. Zhang, X. He, Q. Jiao, Q. Zhang, and J. Han, "AMNet: Learning to align multi-modality for RGB-T tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 8, pp. 7386–7400, Aug. 2024.

[27] C. Li, G. Wang, Y. Ma, A. Zheng, B. Luo, and J. Tang, "A unified RGB-T saliency detection benchmark: Dataset, baselines, analysis and a novel approach," 2017, *arXiv:1701.02829*.

[28] Z. Tu, T. Xia, C. Li, Y. Lu, and J. Tang, "M3S-NIR: Multi-modal multi-scale noise-insensitive ranking for RGB-T saliency detection," in *Proc. MIPR*, Mar. 2019, pp. 141–146.

[29] X. Wang, S. Li, C. Chen, Y. Fang, A. Hao, and H. Qin, "Data-level recombination and lightweight fusion scheme for RGB-D salient object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 458–471, 2021.

[30] Z. Xie et al., "Cross-modality double bidirectional interaction and fusion network for RGB-T salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 8, pp. 4149–4163, Aug. 2023.

[31] H. Zhou, C. Tian, Z. Zhang, C. Li, Y. Xie, and Z. Li, "Frequency-aware feature aggregation network with dual-task consistency for RGB-T salient object detection," *Pattern Recognit.*, vol. 146, Feb. 2024, Art. no. 110043.

[32] 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. 15013–15022.

[33] Y. Cai, H. Bian, J. Lin, H. Wang, R. Timofte, and Y. Zhang, "Retinexformer: One-stage Retinex-based transformer for low-light image enhancement," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 12504–12513.

[34] Q. Yan et al., "HVI: A new color space for low-light image enhancement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2025, pp. 5678–5687.

[35] C. Wei, W. Wang, W. Yang, and J. Liu, "Deep retinex decomposition for low-light enhancement," 2018, *arXiv:1808.04560*.

[36] W. Yang, W. Wang, H. Huang, S. Wang, and J. Liu, "Sparse gradient regularized deep retinex network for robust low-light image enhancement," *IEEE Trans. Image Process.*, vol. 30, pp. 2072–2086, 2021.

[37] C. Chen, Q. Chen, M. Do, and V. Koltun, "Seeing motion in the dark," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3184–3193.

[38] C. Chen, Q. Chen, J. Xu, and V. Koltun, "Learning to see in the dark," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 3291–3300.

[39] 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.

[40] 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)*, 2021, pp. 13567–13576.

[41] 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. (CVPR)*, Miami, FL, USA, 2009, pp. 248–255.

[42] Z. Shu, Y. Luo, Y. Huang, C. Mao, and Z. Yu, "View-interactive attention information alignment-guided fusion for incomplete multi-view clustering," *Expert Syst. Appl.*, vol. 252, Oct. 2024, Art. no. 124258.

[43] R. Hong, X.-P. Chen, Y. Zhou, H. Liu, and T. Wan, "Prototype imputation guided incomplete multi-view clustering," *IEEE Signal Process. Lett.*, vol. 32, pp. 1565–1569, 2025.

[44] B. Jiang et al., "Collaborative similarity fusion and consistency recovery for incomplete multi-view clustering," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2025, vol. 39, no. 17, pp. 17617–17625.

[45] J. Wen et al., "Diffusion-based missing-view generation with the application on incomplete multi-view clustering," in *Proc. ICML*, 2024, pp. 52762–52778.

[46] Y. Zhang et al., "Incomplete multi-view clustering via diffusion contrastive generation," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2025, vol. 39, no. 21, pp. 22650–22658.

[47] J. Pu et al., "Adaptive feature imputation with latent graph for deep incomplete multi-view clustering," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2024, vol. 38, no. 13, pp. 14633–14641.

[48] L. Du, Y. Shi, Y. Chen, P. Zhou, and Y. Qian, "Fast and scalable incomplete multi-view clustering with duality optimal graph filtering," in *Proc. 32nd ACM Int. Conf. Multimedia*, Oct. 2024, pp. 8893–8902.

[49] 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.

[50] Ultralytics. (2025). *yolov5: YOLOv5 in PyTorch*. [Online]. Available: https://github.com/ultralytics/yolov5

[51] 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. 2223–2232.

[52] W. Li, R. Zhao, T. Xiao, and X. Wang, "DeepReID: Deep filter pairing neural network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 152–159.

[53] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer GAN to bridge domain gap for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 79–88.

[54] L. Zheng et al., "MARS: A video benchmark for large-scale person re-identification," in *Proc. ECCV*, 2016, pp. 868–884.

[55] 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.

[56] L. Pang, Y. Wang, Y.-Z. Song, T. Huang, and Y. Tian, "Cross-domain adversarial feature learning for sketch re-identification," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 609–617.

[57] M. Munaro, A. Fossati, A. Basso, E. Menegatti, and L. V. Gool, "One-shot person re-identification with a consumer depth camera," in *Person Re-Identification*. London, U.K.: Springer, 2014, pp. 161–181.

[58] I. B. Barbosa, M. Cristani, A. D. Bue, L. Bazzani, and V. Murino, "Re-identification with RGB-D sensors," in *Proc. ECCV*, 2012, pp. 433–442.

[59] D. Baltieri, R. Vezzani, and R. Cucchiara, "3DPeS: 3D people dataset for surveillance and forensics," in *Proc. Joint ACM Workshop Human Gesture Behav. Understand.*, Dec. 2011, pp. 59–64.

[60] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, "Custom pictorial structures for re-identification," in *Proc. Procedings Brit. Mach. Vis. Conf. (BMVC)*, 2011, pp. 68.1–68.11.

[61] 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.

[62] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 4510–4520.

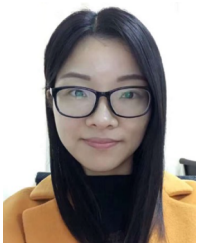
[63] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[64] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 6848–6856.
[65] 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)*, 2018, pp. 480–496.
[66] X. Chang, T. M. Hospedales, and T. Xiang, "Multi-level factorisation net for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 2109–2118.
[67] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Omni-scale feature learning for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3701–3711.
[68] 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. 1–9.
[69] M. Ye, J. Shen, D. Crandall, L. Shao, and J. Luo, "Dynamic dual-attentive aggregation learning for visible-infrared person re-identification," in *Proc. ECCV*, 2020, pp. 229–247.
[70] H. Park, S. Lee, J. Lee, and B. Ham, "Learning by aligning: Visible-infrared person re-identification using cross-modal correspondences," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 12026–12035.
[71] X. Zheng, X. Chen, and X. Lu, "Visible-infrared person re-identification via partially interactive collaboration," *IEEE Trans. Image Process.*, vol. 31, pp. 6951–6963, 2022.
[72] K. Ren and L. Zhang, "Implicit discriminative knowledge learning for visible-infrared person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 393–402.
[73] J. Li, Q. Zhen, Y. Yang, Y. Li, Z. Dong, and C. Yang, "Prototype-driven multi-feature generation for visible-infrared person re-identification," in *Proc. ICASSP - IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2025, pp. 1–5.
[74] Y. Feng et al., "Homogeneous and heterogeneous relational graph for visible-infrared person re-identification," *Pattern Recognit.*, vol. 158, Feb. 2025, Art. no. 110981.
[75] X. Yu, N. Dong, L. Zhu, H. Peng, and D. Tao, "CLIP-driven semantic discovery network for visible-infrared person re-identification," *IEEE Trans. Multimedia*, vol. 27, pp. 4137–4150, 2025.
[76] C. Wang et al., "Cross-modal pattern-propagation for RGB-T tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7064–7073.

Zi Wang received the B.Eng. and Ph.D. degrees from the School of Computer Science and Technology, Anhui University. He is currently with the School of Biomedical Engineering, Anhui Medical University. He is primarily engaged in research on computer vision, medical image processing, and multi-modal learning.

Chenglong Li (Member, IEEE) received the M.S. and Ph.D. degrees from the School of Computer Science and Technology, Anhui University, Hefei, China, in 2013 and 2016, respectively. From 2014 to 2015, he was a Visiting Student with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China. He was a Post-Doctoral Research Fellow with the Center for Research on Intelligent Perception and Computing, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. He is currently a Professor with the School of Artificial Intelligence, Anhui University. His research interests include computer vision and deep learning. He was a recipient of the ACM Hefei Doctoral Dissertation Award in 2016.

Pengyu Li received the B.Eng. degree in 2021. He is currently pursuing the M.Eng. degree with the School of Artificial Intelligence, Anhui University, Hefei, China. His research interests include computer vision, multi-modal intelligence, and object re-identification.



Aihua Zheng received the B.Eng. and M.Eng. degrees in computer science and technology from Anhui University of China in 2006 and 2008, respectively, and the Ph.D. degree in computer science from the University of Greenwich, U.K., in 2012. She visited the University of Stirling and Texas State University in 2013 and 2019, respectively. She is currently a Full Professor and a Ph.D. Supervisor with the School of Artificial Intelligence, Anhui University. Her research interests include vision-based artificial intelligence and pattern recognition. Especially on object re-identification, audio visual computing, and multi-modal intelligence.

Jin Tang received the B.Eng. degree in automation and the Ph.D. degree in computer science from Anhui University, Hefei, China, in 1999 and 2007, respectively. He is currently a Professor and a Ph.D. Supervisor with the School of Computer Science and Technology, Anhui University. His research interests include computer vision, pattern recognition, and machine learning.

Bin Luo received the B.Eng. degree in electronics and the M.Eng. degree in computer science from Anhui University of China, Hefei, China, in 1984 and 1991, respectively, and the Ph.D. degree in computer science from the University of York, York, U.K., in 2002. He is currently a Professor with Anhui University, China. He has authored or co-authored more than 200 papers in journals and refereed conferences. His research interests include random graph-based pattern recognition, image and graph matching, and spectral analysis. He is also the Chair of the IEEE Hefei Subsection. He was a Peer Reviewer of international academic journals, such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, *Pattern Recognition*, and *Pattern Recognition Letters*.