Occluded Visible-Infrared Person Re-Identification

Yujian Feng, Yimu Ji, Fei Wu, Guangwei Gao, *Senior Member, IEEE*, Yang Gao, Tianliang Liu, Shangdong Liu, Xiao-Yuan Jing, and Jiebo Luo, *Fellow, IEEE*

***Abstract*—Visible-infrared person re-identification (VI-ReID) aims to match person images between the visible and near-infrared modalities. Previous VI-ReID methods are based on holistic pedestrian images and achieve excellent performance. However, in real-world scenarios, images captured by visible and near-infrared cameras usually contain occlusions. The performance of these methods degrades significantly due to the loss of information of discriminative features from the occlusion of the images. We define visible-infrared person re-identification in this occlusion scene as Occluded VI-ReID, where only partial content information of pedestrian images can be used to match images of different modalities from different cameras. In this paper, we propose a matching framework for occlusion scenes, which contains a local feature enhance module (LFEM) and a modality information fusion module (MIFM). LFEM adopts Transformer to learn features of each modality, and adjusts the importance of patches to enhance the representation ability of local features of the non-occluded areas. MIFM utilizes a co-attention mechanism to infer the correlation between each image for reducing the difference between modalities. We construct two occluded VI-ReID datasets, namely Occluded-SYSU-MM01 and Occluded-RegDB datasets. Our approach outperforms existing state-of-the-art methods on two occlusion datasets, while remains top performance on two holistic datasets.**

***Index Terms*—Occluded Visible-infrared Person Re-identification, Visual Transformer, Co-attention mechanism, Occluded VI-ReID datasets.**

I. Introduction

Visible-infrared person re-identification (VI-ReID) [1], [2], [3], [4] aims to identify the same pedestrian captured by visible light and near-infrared cameras under complex scenarios. To match the identity information of a person across modalities, one has to deal with large appearance changes of the person caused by a variety of condition changes including lighting, view angle, and pose. Generally, previous VI-ReID methods [5], [6], [7], [8], [9] used global features or local features of the holistic image as shown in Fig. 1(a) to extract effective feature representations and achieved satisfactory performance.

In real-world scenarios, occlusion is one of the most challenging problems to solve because the loss of information from an image is irreversible. As shown in Fig. 1(b), occlusion is common in crowded public spaces with cluttered backgrounds and may be caused by obstacles such as other people in the scene, car, and chair. When a person is partially occluded, the representations extracted from the whole image might involve ambiguous noise information from the occluded regions. This phenomenon might lead to wrong retrieval results in the model, because the model does not differentiate the occluded and person regions.Traditional occluded ReID methods focus on matching occluded images in visible scenes and achieving nice performance, which cannot effectively identify occluded pedestrians in near-infrared scenes. Therefore, the problem of occluded pedestrian matching between the two modalities needs to be well studied. We call this scenario **Occluded VI-ReID**.

Occluded VI-ReID is widely present in the field of police reconnaissance. For example, in case of an investigation, criminals often commit crimes at night and obscure their bodies through obstacles, so the cameras often capture near-infrared occluded images of criminals. The police need to obtain the identity information of pedestrians from the gallery (visible light images/ near-infrared images) by matching the probe (near-infrared images/visible light images) of pedestrians. However, the probe images captured by the camera is occluded as shown in Fig. 1(c). The police gallery database includes non-occluded visible and near-infrared images of pedestrians as shown in Fig. 1(d). Therefore, it is necessary to study the Occluded VI ReID to improve the pedestrian matching effect under the two modalities.

Manuscript received 21 July 2022; revised 20 October 2022; accepted 30 November 2022. Date of publication 16 December 2022; date of current version 8 May 2023. This work was supported in part by the National Natural Science Foundation of China under Grants 61902194, 62076139, 62176069, 61933013, and 61972212, in part by the National Key Research and Development Program of China under Grant 2018AAA0103302, in part by the Natural Science Foundation of Jiangsu Province (Higher Education Institutions) under Grants BK20170900, 19KJB520046, and 20KJA520001, and BK20190089, in part by the 14th Five-Year Plan Project of Equipment Development Department under Grant 315107402, in part by the Innovative and Entrepreneurial talents projects of Jiangsu Province, in part by the Jiangsu Planned Projects for Postdoctoral Research Funds under Grant 2019K024, in part by the Postgraduate Research and Practice Innovation Program of Jiangsu Province under Grants KYCX19_0921 and KYCX19_0906, in part by the Open Research Project of Zhejiang Lab under Grant 2021KF0AB05, in part by Future Network Scientific Research Fund Project under Grant FNSRFP-2021-YB-15, and in part by the 1311 Talent Program of Nanjing University of Posts and Telecommunications. The guest editor coordinating the review of this manuscript and approving it for publication was Dr. Cairong Zhao. *(Corresponding authors: Yimu Ji; Fei Wu; Guangwei Gao.)*

Yujian Feng is with the School of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing 210003, China (e-mail: fengyujian_904@163.com).

Yimu Ji and Shangdong Liu are with the School of Computer Science and Technology, NJUPT, Nanjing 210003, China (e-mail: jiym@njupt.edu.cn; lsd@njupt.edu.cn).

Fei Wu, Guangwei Gao, and Yang Gao are with the College of Automation, NJUPT, Nanjing 210023, China (e-mail: wufei_8888@126.com; csggao@gmail.com; gy731222@163.com).

Tianliang Liu is with the School of Communication and Information Engineering, NJUPT, Nanjing 210003, China (e-mail: liutl@njupt.edu.cn).

Xiao-Yuan Jing is with the Guangdong Provincial Key Laboratory of Petrochemical Equipment Fault Diagnosis, Guangdong University of Petrochemical Technology, Maoming 525000, China, and also with the School of Computer, Wuhan University, Wuhan 430072, China (e-mail: jingxy_2000@126.com).

Jiebo Luo is with the Department of Computer Science, University of Rochester, Rochester, NY 14627-0001 USA (e-mail: jiebo.luo@gmail.com).

Digital Object Identifier 10.1109/TMM.2022.3229969

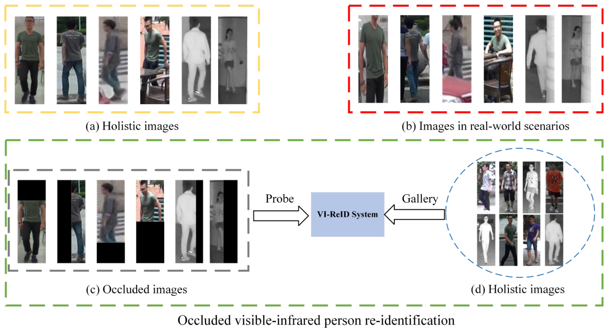

(a) Holistic images
(b) Images in real-world scenarios
Probe
VI-ReID System
Gallery
(c) Occluded images
(d) Holistic images
Occluded visible-infrared person re-identification

Fig. 1. The Occluded VI-ReID scenario where probe images are occluded and all gallery images are holistic.

TABLE I
THE RESULTS OF STATE-OF-THE-ARTS ON THE SYSU-MM01 AND OCCLUDED-SYSU-MM01 DATASETS WHERE R DENOTES RANK. THE MAP DENOTES MEAN AVERAGE PRECISION SCORE

Method	SYSU-MM01				Occluded-SYSU-MM01			
	All-Search		Indoor-Search		All-Search		Indoor-Search	
	R1	mAP	R1	mAP	R1	mAP	R1	mAP
AGW [10]	47.50	47.65	54.17	62.97	18.20	20.88	22.24	32.27
CM-NAS [11]	61.99	60.02	62.14	66.75	34.47	34.85	34.01	43.34
CAJ [12]	69.88	66.89	76.26	80.37	31.50	31.72	36.96	46.92

There are two challenges in the Occluded VI-ReID scenario, as follows: 1) since the pedestrian image only contains partial information, e.g., the pedestrian image can only be observed in the lower body position and the more obvious parts of the upper body are occluded, it is difficult to distinguish the true identity of the pedestrian due to the high similarity of the lower body posture to that of others. 2) similar to VI-ReID need to mitigate modality differences, Occluded VI-ReID also needs to reduce the difference between visible light images and near-infrared images captured by different cameras to improve the performance of the model. In addition, compared to VI-ReID, in Occluded VI-ReID, the modality information of the image is missing due to occluded regions, which makes the difference between modalities become irregular and asymmetric, i.e., the size and location of the modality information in the non-occluded region is inconsistent. Table I lists the results of three representative VI-ReID methods (AGW [10], CM-NAS [11] and CAJ [12]) on the non-occluded dataset (SYSU-MM01 [3]) and occluded dataset (Occluded-SYSU-MM01). Compared with the performance on the SYSU-MM01 dataset, we can observe that the performance of these methods decreases significantly on Rank-1 (R1) and mAP in the occluded dataset. Details about experimental results will be introduced in Section IV.

To tackle these two challenges in Occluded VI-ReID, we propose two strategies to enhance the Occluded VI-ReID performance, including: (1) In the feature extraction stage, the model should pay more attention to the non-occluded part and obtain saliency representations of local features. (2) Facing the modality gap, the model tries to reduce modality differences by mining intra-modal feature representations and exchanging inter-modal information. In this paper, based on two strategies, we propose a matching framework consisting of two parts, local feature enhance module (LFEM) and modality information fusion module (MIFM).

Specifically, LFEM adopts Transformer [13], [14], [15] to extract patches, regards patches as local features and the set of all patches are regarded as global feature. To focus on the non-occluded areas, LFEM first adjusts the importance of each local feature by generating an attention weight based on the similarity between each local feature and the global feature. Then, LFEM learns static contextual representations of local features and obtains dynamic attention matrix, correlating static and dynamic contextual representations to obtain saliency representations of local features. Therefore, each local feature can autonomously generate a contextual feature representation, which enhances the saliency of non-occluded semantic regions and compresses noise and interference in occluded regions.

Meanwhile, MIFM has been proposed to alleviate the difference between the visible and near-infrared modalities. MIFM first constructs the dense intra-modality interaction mechanisms (visible-to-visible or infrared-to-infrared) by multi-head self-attention mechanism [13], [14], [15] to mine intra-modality feature representations. Then, we utilize an inter-modality interaction mechanism that uses a single token of each modality as a query to exchange information with other modalities to fuse the information from the features between the modalities. In this way, the ability to fuse the feature representations of the two modalities is improved by taking into account the dense symmetric interaction between the images of the two modalities.

The main contribution of this paper can be summarized as follows:

1) To the best of our knowledge, we first focus on Occluded VI-ReID and introduce a matching framework including the local feature enhance module (LFEM) and the modality information fusion module (MIFM) for occlusion scenes.
2) To simulate the occlusion environment in the real-world, we propose two occluded datasets (Occluded-SYSU-MM01 and Occluded-RegDB), and experimental results demonstrate that our approach outperforms the state-of-the-arts by a large margin.

II. RELATED WORKS

A. *Visible-Infrared Person Re-Identification*

Visible-infrared person re-identification (VI-ReID) is designed to match a query of one modality with a gallery set of another modality, i.e., visible modality and near-infrared modality. The two main difficulties in VI-ReID are the inter-modality differences in images due to the different imaging mechanisms of visible and near-infrared cameras, and the intra-modality variations resulting from the camera viewing changes.

To tackle a challenge caused by modality discrepancy and alleviate the intra-modality variations, Ye et al. [5], [6] proposed global feature learning to handle inter-modality difference and intra-modality variation, enhance the discriminative power of features. Lu et al. [8] explored the potential of both the modality-shared information and the modality-specific characteristics to improve the performance. Wu et al. [16] discovered cross-modality nuances in different patterns which introduced a modality alleviation module and a pattern alignment module. Wu et al. [17] exploited same-modality similarity as the constraint to guide the learning of cross-modality similarity. Chen et al. [18] utilized the structural and positional information to learn semantic-aware sharable modality features. Zhao et al. [19] learned the color-irrelevant features and aligns the identity-level feature distributions. Park et al. [20] encouraged pixel-wise associations between cross-modality local features to facilitate discriminative feature learning. Ye et al. proposed a strong baseline AGW method [10] which utilized the attention mechanism to extract discriminative part-aggregated features. Fu et al. designed the CM-NAS method [11] to study the manually designed architectures, and identify that appropriately separated Batch Normalization layers. Ye et al. proposed the CAJ method [12] to simultaneously handle the intra and cross-modality variations by enhancing channel-mixed learning strategy.

However, these previous methods did not consider a situation where various obstructions occlude the target person e.g., tree, cars, railing, and another person in visible and near-infrared cameras. As shown in Table I, we observed a significant decrease in the results of these methods. In this paper, we first focus on Occluded VI-ReID task and propose an effective approach for solving the retrieval problem when pedestrians are occluded.

B. Occluded Person Re-Identification

In real-world scenarios, occlusion occurs when only partial regions of the target person are available for person reidentification (ReID). Occluded person ReID aims to match the occluded probe image with the holistic image of the gallery [21], [22], [23], [24], [25]. In recent years, in the visible light modality, there have been some methods to deal with the occluded person re-identification problem. Zheng et al. [26] proposed a local patch-level matching mode and a global part-based matching to address the partial ReID problem. He et al. [27] exploited dictionary learning to calculate the similarity between different spatial feature maps. He et al. [28] learned correspondence between image patches without any additional part-level supervision. Gao et al. [29] employed a texture alignment scheme with semantic visibility and designed a human pose-based partial region alignment scheme to solve the occlusion problem. Zhuo et al. [30] designed a co-saliency network and a cross-domain simulator to highlight meaningful parts. Miao et al. [31] introduced pose-guided feature alignment to disentangle the useful information from the occlusion noise. Jia et al. [32] measured image similarity by automatically disentangling the representations of undefined semantic components. Li et al. [33] utilized Transformer encoder-decoder architecture to diverse part discovery.

Although these methods are excellent for the task of pedestrian image occlusion in visible scenarios, they are designed for single-modality scenes and can not be directly generalized to cross-modality scenarios. In this paper, we focus on Occluded VI-ReID task and propose a matching framework to address the pedestrian re-identification problem in visible and near-infrared scenarios.

C. Visual Transformer

Transformer structure is initially proposed to handle sequential data the field of natural language processing in [13]. Recently, many studies also show its effectiveness for vision tasks [14], [34], [35], [36], [37], [38], [39]. ViT [14] adopted pure Transformer and achieves excellent performance in image classification. IPT [35] toke advantage of transformers to achieve state-of-the-art performance by using large-scale pre-training on a number of image processing tasks. Bert [34] is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. Li et al. [36] utilized contextual information between input keys to guide the learning of dynamic attention matrices to enhance visual representations. In this paper, we utilize ViT [14] structure to preserve semantic and detail information from a global view for each single modality in Occluded VI-ReID task.

UniT [40] model encodes each input modality with an encoder and makes predictions on each task with a shared decoder over the encoded input representations. [41] proposed framework, which separates the person picture into several regions to match the linguistic referring words with visual-language co-attention. [42] designed a discriminative triad through which a query can be converted into one or multiple discriminative triads in a very scalable way. Compared to [40], [41], [42], our approach constructs the dense intra-modality interaction mechanisms to mine intra-modality feature representations, and utilizes an inter-modality interaction mechanism to fuse the information from the features between the modalities.

III. Our Approach

In this section, we introduce the proposed framework for Occluded VI-ReID. Firstly, we give a brief overview of the network architecture. Secondly, we present local feature enhance module (LFEM) to enhance the saliency representation of local features for each modality. The modality information fusion module (MIFM) is introduced to reduce difference between modalities. Finally, we propose a total loss function for optimizing the parameters of our framework.

A. Overview

The overview of our approach in presented as shown in Fig. 2. Firstly, visible and near-infrared images are fed into a two-stream feature extraction module consisting of ResNet50 [43] pre-trained by ImageNet [44] to extract features, obtaining visible features X_{cnn}^{vis} and near-infrared features X_{cnn}^{inf}. X_{cnn}^{vis} and X_{cnn}^{inf} are fed to the transformer encoder, and $X^{vis} = \left\{x_1^{vis}, x_2^{vis}, \ldots, x_n^{vis}\right\}$ and $X^{inf} = \left\{x_1^{inf}, x_2^{inf}, \ldots, x_n^{inf}\right\}$ are obtained. n is the number of patches in an image. Secondly, the X^{vis} and X^{inf} are fed into local feature enhance module (LFEM) to obtain saliency enhancement feature representations $\bar{X}_{vis}$ and $\bar{X}_{inf}$. Meanwhile, modality information fusion module (MIFM) utilizes X^{vis} and X^{inf} to construct intra-modality

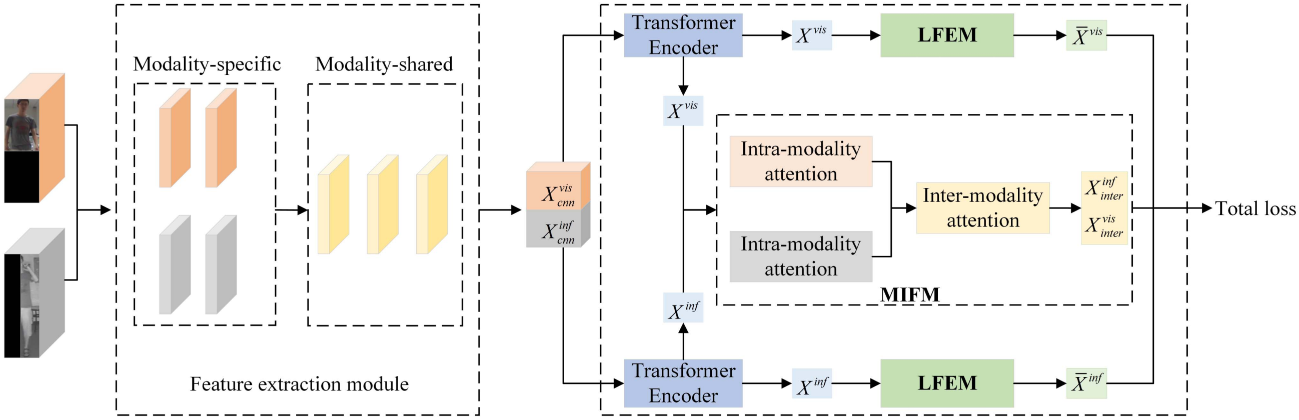


Fig. 2. Illustration of our proposed framework, which contains four components: feature extraction module, local feature enhance module (LFEM), modality information fusion module (MIFM) and total loss.

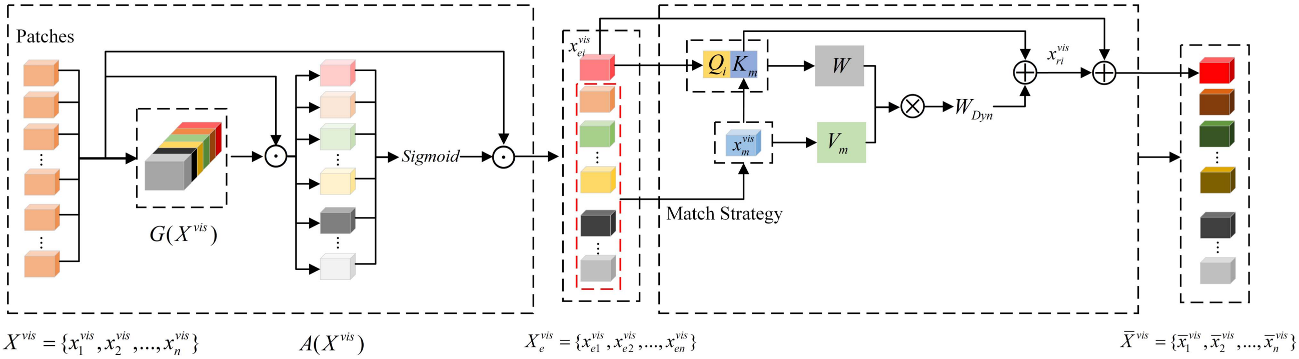

Patches
$G(X^{vis})$
Sigmoid
x^{vis}_{ei}
Q_i
K_m
W
x^{vis}_m
V_m
W_{Dyn}
x^{vis}_{ri}
Match Strategy
$X^{vis} = \{x_1^{vis}, x_2^{vis}, ..., x_n^{vis}\}$
$A(X^{vis})$
$X_e^{vis} = \{x_{e1}^{vis}, x_{e2}^{vis}, ..., x_{en}^{vis}\}$
$\bar{X}^{vis} = \{\bar{x}_1^{vis}, \bar{x}_2^{vis}, ..., \bar{x}_n^{vis}\}$

Fig. 3. Illustration of our proposed local feature enhance module (LFEM) for visible modality, which enhances the saliency representation of local features from non-occluded regions. The same process is adopted for infrared modality.

and inter-modality interactions to obtain X^{vis}_{inter} and X^{inf}_{inter}. Finally, features from LFEM and MIFM outputs are aggregated together and fed into the total loss function.

B. Local Feature Enhance Module

As shown in Fig. 3, we first define and calculate the feature strength of each patch by utilizing the information from all patches, then select the patch in the non-occluded area through the designed matching strategy, and finally fuse the information of multiple patches to obtain discriminative local features.

Firstly, for visible modality images, given the feature patches $X^{vis} = \{x^{vis}_1, x^{vis}_2, \ldots, x^{vis}_n\}$ by Transformer encoder. We regard patches as local features and the set of all patches are regarded as global feature. Since the image global features are not dominated by occlusion noise, i.e., the feature strength of non-occluded regions is stronger than that of occluded regions, we approximate the global statistical feature strength $G(X^{vis})$ with an average function as follows:

$$G\left(X^{vis}\right) = \frac{1}{n}\sum_{i=1}^{n} x^{vis}_i \tag{1}$$

where n is the number of patches in an image. x^{vis}_i represents the feature patch. Then, to learn the corresponding strength coefficient for each patch, we generate a weight matrix $A(X^{vis})$ by computing the similarity of each patch x^{vis}_i and global feature $G(X^{vis})$ as follows:

$$A\left(X^{vis}_i\right) = \frac{G\left(X^{vis}\right)\cdot x^{vis}_i - \mu_c}{\sigma_c + \varepsilon} \tag{2}$$

where ε is a constant added for numerical stability. μ_c and σ_c [45], [46] are to prevent the bias magnitude of the coefficients.

$$\mu_c = \frac{1}{n}\sum_{j}^{n} G\left(X^{vis}\right)\cdot x^{vis}_j,$$

$$\sigma_c = \sqrt{\frac{1}{n}\sum_{j}^{n}\left(G\left(X^{vis}\right)\cdot x^{vis}_j - \mu_c\right)} \tag{3}$$

To obtain enhanced local feature x^{vis}_{ei}, x^{vis}_i is scaled by the weight matrix $A(X^{vis})$ via a sigmoid function:

$$x^{vis}_{ei} = x^{vis}_i \cdot sigmoid\left(A\left(x^{vis}_i\right)\right) \tag{4}$$

The set of enhanced local features x^{vis}_{ei} is X^{vis}_e, and $X^{vis}_e = \{x^{vis}_{e1}, x^{vis}_{e2}, \ldots, x^{vis}_{en}\}$. The feature strength response of non-occluded regions is further enhanced by the interaction of the features of each patch with the global features of all patches.

Secondly, we develop a match strategy for enhanced local part features to automatically select non-occluded local features. For each patch x_{ei}^{vis}, as shown in (5), we can find the most similar feature f_m^t in patches $x_{e1}^{vis}, x_{e2}^{vis}, \ldots, x_{et}^{vis}$, t is the number of remaining patches.

$$f_m^t = \max\left(\frac{< x_{ei}, (x_{e1}, x_{e2}, \ldots, x_{et}) >}{\|x_{ei}\| \, \|x_{e1}, x_{e2}, \ldots, x_{et}\|}\right) \tag{5}$$

where $< \cdot, \cdot >$ represents the inner product. Then x_{ei}^{vis} and f_m^t are added to form x_m^{vis} as follows:

$$x_m^{vis} = x_{ei}^{vis} + f_m^t \tag{6}$$

Then, we utilize the contextual information between patches to learn the dynamic attention matrix, which improves the correlation between local features from non-occluded areas. As shown in Fig. 3, after obtaining patch features X_e^{vis}, x_{ei}^{vis} is fed into three linear projections to generate query Q_i, key K_i, and value V_i, where $\sqrt{d}$ is utilized to normalize for numerical stability. In the same way, the query Q_m, key K_m, and value V_m from x_m can been obtained. In (8), we utilize Q_i and K_m to generate attention matrix W.

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d}}\right)V, \tag{7}$$

$$W = W_\alpha W_\beta [K_m, Q_i] \tag{8}$$

where W_α and W_β represent the weight matrices with and without $Relu$ activation function, respectively. $[K_m, Q_i]$ represents concatenate key K_m and query Q_i. Then, we compute the feature map by aggregating V_i according to the contextual attention matrix W to obtain dynamic attention W_{dyn}:

$$W_{dyn} = V_i \otimes W \tag{9}$$

where $\otimes$ denotes the local matrix multiplication operation. By this way, W_{dyn} captures the dynamic feature interaction between patches. The patch relation feature x_{ri}^{vis} is fused between the patch static context information K_m and the dynamic context W_{dyn}. As shown in (10), each feature patch $\bar{x}_i^{vis}$ is the aggregation of x_{ei}^{vis} and x_{ri}^{vis}, and the set of features is $\bar{X}^{vis} = \left\{\bar{x}_1^{vis}, \bar{x}_2^{vis}, \ldots, \bar{x}_n^{vis}\right\}$.

$$\begin{aligned} x_{ri}^{vis} &= K_m + W_{dyn}, \\ \bar{x}_i^{vis} &= x_{ei}^{vis} + x_{ri}^{vis} \end{aligned} \tag{10}$$

We apply a similar process to the infrared modality feature X^{inf} and then obtain $\bar{X}^{inf} = \left\{\bar{x}_1^{inf}, \bar{x}_2^{inf}, \ldots, \bar{x}_n^{inf}\right\}$.

C. Modality Information Fusion Module

The modality information fusion module (MIFM) is designed to construct the relationship between features in visible and near-infrared modalities. We utilize intra-modality attention to model the dense intra-modality interactions (visible-visible and infrared-infrared). Then, inter-modality attention is proposed to model the inter-modality interactions (visible-infrared and infrared-visible).

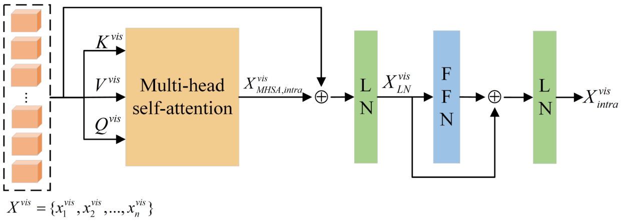


Fig. 4. The detail of the intra-modality attention for visible modality. The similar detail illustrates intra-modality attention for infrared modality.

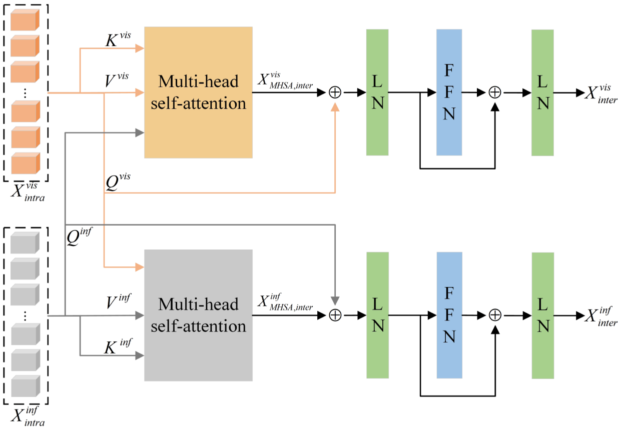


Fig. 5. The detail of the inter-modality attention.

In detail, as shown in Fig. 4, for the visible modality, the intra-modality attention is calculated by the dot product of the query Q^{vis} with all keys K^{vis}, and dividing each by $\sqrt{d}$, then applying a $softmax$ function to obtain the attention weights on the values V^{vis} by (7). To further mine the intra-modality representation capacity of the attended features, the multi-head self-attention (MHSA) is utilized. MHSA includes h paralleled 'heads,' which correspond to the independent scaled dot-product attention function. The attended output features X_{MHSA}^{vis} is given by:

$$\begin{aligned} X_{MHSA,intra}^{vis} &= MHSA\left(Q^{vis}, K^{vis}, V^{vis}\right) \\ &= [head_1, head_2, \ldots, head_h] \, W^o \end{aligned} \tag{11}$$

$$head_h = Attention\left(Q^{vis}W_t^Q, K^{vis}W_t^K, V^{vis}W_t^V\right) \tag{12}$$

where W_t^Q, W_t^K, W_t^V are the projection matrices for t-th head. h is the number of paralleled 'heads'.

Then, the output and input of the MHSA layer are connected by residual connections and a layer normalization (LN) in (13). The feed-forward network (FFN) consisting of two linear projections is applied after the MHSA layer, as shown (14).

$$X_{LN}^{vis} = LN\left(X^{vis} + X_{MHSA,intra}^{vis}\right), \tag{13}$$

$$X_{intra}^{vis} = LN\left(X_{LN}^{vis} + FFN\left(X_{LN}^{vis}\right)\right), \tag{14}$$

A similar process obtains X_{intra}^{inf} of infrared modality. In this way, the intra-modality attention can mine the feature information within modality from a global view.

As shown in Fig. 5, after obtaining the Q^{vis}, K^{vis}, V^{vis} and Q^{inf}, K^{inf}, V^{inf} of the visible and infrared modalities from

(7), we utilize the inter-modality attention mechanism to interact with the feature information of the two modalities. The inter-modality attention takes two groups of input features X_{vis} and X_{inf}, where X_{vis} guides the attention for X_{inf} and vice versa, and models the pairwise relationship between each paired samples $< x_i^{vis}, x_i^{inf} >$ from X_{vis} and X_{inf}. Specifically, the $K^{vis}, V^{vis}, Q^{vis}$ from visible modality and the the K^{inf}, V^{inf}, Q^{inf} from infrared modality are fed into MHSA. Then output feature $X_{MHSA,inter}^{vis}$ and $X_{MHSA,inter}^{inf}$ can be denoted as follows:

$$\begin{aligned} X_{MHSA,inter}^{vis} &= MHSA\left(Q^{inf}, K^{vis}, V^{vis}\right), \\ X_{MHSA,inter}^{inf} &= MHSA\left(Q^{vis}, K^{inf}, V^{inf}\right) \end{aligned} \tag{15}$$

Similar to intra-modality attention, for $X_{MHSA,inter}^{vis}$ and $X_{MHSA,inter}^{inf}$, the next process is to adopt strategies (13) and (14). The finally output results X_{inter}^{vis} and X_{inter}^{inf} of inter-modality attention can be obtained through interacting the information of two modalities.

D. Total Loss

To learn discriminative features, we combine the identity loss L_{id} and triplet center loss L_{tc} [47], [48], [49] as our learning objective to simultaneously deal with both cross-modality discrepancy and intra-modality variations in the common feature space.

The L_{id} aims to learn the discriminative feature representations by utilizing modality-specific information, which encourages an identity-invariant feature representation and distinguishes different persons within modality. The L_{id} is denoted by:

$$L_{id} = -\sum_{i=1}^{N} y_i \log(p_i) \tag{16}$$

where N is the number of identities. y_i is the predicted probability, and p_i is the groundtruth probability.

The L_{tc} learns two intra-modality centers (c_i^{vis}, c_i^{inf}) for each class and requires samples from the same class to be closer to the center than samples from different classes. The triplet-center loss is as follows:

$$\begin{aligned} L_{tc} = &\sum_{i=1}^{P}\left[\rho + D\left(c_i^{vis}, c_i^{inf}\right) - \min_{j\neq i} D\left(c_i^{vis}, c_j^{vis}\right)\right] \\ &+ \sum_{i=1}^{P}\left[\rho + D\left(c_i^{inf}, c_i^{vis}\right) - \min_{j\neq i} D\left(c_i^{inf}, c_j^{inf}\right)\right] \end{aligned} \tag{17}$$

where P is the number of identities in a mini-batch. $c_i^{vis} = \frac{1}{M}\sum_{j=1}^{M} x_{i,j}^{vis}, c_i^{inf} = \frac{1}{M}\sum_{j=1}^{M} x_{i,j}^{inf}$, M denotes the number of single modality images. $x_{i,j}^{vis}$ and $x_{i,j}^{inf}$ represent the j-th features of visible and infrared modality in the i-th identity.

The L_{tc} handles both the intra-class and inter-class variations simultaneously on visible and infrared modalities in the common feature space, and further ensures the intra-class feature compactness and enhances the distinguishability of inter-class features.

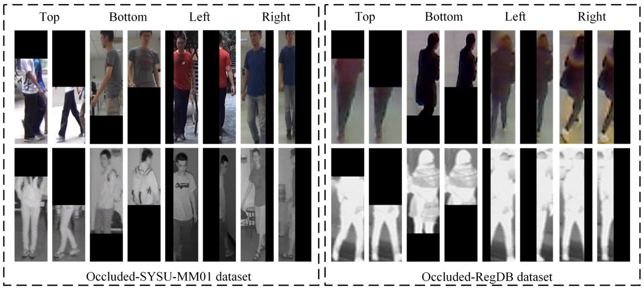


Fig. 6. Example images on the Occluded-SYSU-MM01 and Occluded-RegDB dataset. The occluded position of the image is random up, down, left, and right, and the size of the occluded area (black block) is one-quarter or one-half.

The training process of our approach is shown in Algorithm 1. Before the training stage, the occluded datasets are constructed. During the training stage, given a mini-batch of images, we utilized LFEM and MIFM to obtaining the discriminative features and aggregate these features. The entire feature extractor containing LFEM and MIFM are trained together with the total loss. The total loss function can be represented as:

$$L_{total} = \alpha L_{id} + \beta L_{tc} \tag{18}$$

where α and β are the balance parameters.

IV. Occluded Datasets

To facilitate the research on the Occluded VI-ReID, as shown in Fig. 6, we introduce two datasets named **Occluded-SYSU-MM01** and **Occluded-RegDB**, derived from SYSU-MM01 [3] and RegDB [4].

To construct the occluded dataset, we refer to these single modality methods containing [21], [26], [50], [51], [52], [53], [54] in the single modality. [26] need a manually cropping operation to create partial person re-identification datasets. [50], [51] adopted the rectangle region by different colors to simulate occluded person, and [21], [54] utilized a black patch to generate occluded person. In addition, according to [53], in various real-world scenarios, common occlusions can be categorized into four locations (top, bottom, left, right) and two areas (half, quarter).

Following these methods, we simulate the occlusion environment in the real world on the visible light images and near-infrared images. In detail, half of the images of all pedestrians in the training set are occluded, all query images are occluded, and all gallery images are non-occluded. We use the black patch as the color of the occluded area. The occlusions can be categorized into four positions (top, bottom, left, right) and two areas (half, quarter), the position and area occluded in each image are random.

As shown in Table II, Occluded-SYSU-MM01 dataset consists of 30,071 visible images (VISs) and 15,792 near-infrared images (NIRs) of 491 identities. The training set contains 395 identities, 22,258 VISs and 11,909 NIRs. The test set contains 96 identities, of which 3803 NIRs are used for the query and 301

Algorithm 1: The training procedure of our approach.

```
Input: Training set: $X_{train}$
Output: Optimized model parameters $\Theta_1$, $\Theta_2$, $\Theta_3$, $\Theta_4$
Obtain $X_{train}$ from occluded VI-ReID datasets.
Pretrain the ResNet50 on ImageNet with parameters $\Theta_1$, initialize the Transformer encoder parameters $\Theta_2$, LFEM parameters $\Theta_3$ and MIFM parameters $\Theta_4$.
for each mini-batch $B \subset X_{train}$ do
    Obtain $\bar{X}^{vis}$ and $\bar{X}^{inf}$ by LFEM. // (1)–(10)
    Obtain $X^{vis}_{inter}$ and $X^{inf}_{inter}$ by MIFM. // (11)–(15)
    Aggregate $\bar{X}^{vis}$, $\bar{X}^{inf}$, $X^{vis}_{inter}$ and $X^{inf}_{inter}$.
    Calculate $L_{id}$, $L_{tc}$ by (16), (17).
    Optimize parameters $\Theta_1$, $\Theta_2$, $\Theta_3$, $\Theta_4$ according to (18).
end for
```

randomly selected VISs are used as the gallery. The images are captured by 4 visible light cameras and 2 near infrared cameras. In all-search mode, all images are captured by four visible cameras. In indoor-search mode, the gallery set only contains two indoor visible cameras. The occlusion positions of the image are randomly occluded up, down, left and right, and the size of the occlusion area is 1/2 or 1/4.

As shown in Table II, Occluded-RegDB dataset contains 412 identities and each identity has 10 VISs and NIRs. The training set includes randomly selecting 206 identities (with 2,060 images), and the test set contains the rest 206 identities (with 2,060 images). The images are captured by a visible light camera and a near-infrared camera. It has two different retrieval settings, including both visible-infrared and infrared-visible retrieval performance. The occlusion position of the image and the size of the occlusion area are random, top and bottom, left, right, and 1/2 or 1/4 respectively.

V. Experiments

A. Datasets

1) Holistic Datasets: SYSU-MM01 dataset [3] captures pedestrian images through four visible cameras and two near-infrared cameras. It contains 30,071 visible images and 15,792 NIR images of 491 identities. 22,258 visible images of 395 identities and 11,909 NIR images composed the training set. In the test set, 3803 NIR images were used for querying and 301 randomly selected visible images were used as gallery.

RegDB [4] dataset is constructed by dual-camera (one visible and one near-infrared camera) systems, and includes 412 persons. Each person has 10 visible light images and NIR images. The randomly selected 206 identities are used in the training stage and the remaining equivalent for the testing stage.

2) Occluded Datasets: Details of Occluded-SYSU-MM01 and Occluded-RegDB datasets are presented in Section IV.

We conduct experiments on four VI Re-ID datasets including Occluded-SYSU-MM01 and Occluded-RegDB, SYSU-MM01 and RegDB, to evaluate the performance of our approach.

B. Experimental Settings

1) Evaluation Metrics: The Cumulative Matching Characteristic (CMC) and mean average precision (mAP) are used as evaluation metrics. CMC measures the probability that the correct image appears in the top-r retrieval results. Similar to the experiment setting of [9], [10], the results of Ocluded-SYSU-MM01 are evaluated with *all-search* mode and *indoor-search* mode. On the Occluded-RegDB, two evaluation ways are used, i.e., visible-search-infrared (V-I) and infrared-search-visible (I-V), and the performance is averaged over ten trials on random training/testing splits.

2) Implementation Details: Our approach is implemented with PyTorch and trained on an NVIDIA 3090 GPU. Following the existing VI-ReID methods [5], [9], [10], [49], the ResNet50 [43] model pre-trained on ImageNet [44] is adopted as our CNN backbone network, which includes modality-specific networks for different modalities, and the remaining blocks are used as the modality-shared network. The input images are resized to 288 × 144. The initial learning rate is 0.1. The stochastic gradient descent (SGD) optimizer is adopted, and we set the momentum parameter to 0.9. The ViT [14] is used as our Transformer backbone network, the patches are generated with 3 × 1, and the number of heads is set to 128. The margin value ρ is set to 0.5.

In this paper, we propose a new baseline that combines the advantages of ResNet50 and Transformer for better performance. The CNN-based VI-ReID approach is concerned with the ability to capture details of person images by considering local spatial contexts of different complexity, but is weaker in global feature modeling. The Pure-Transformer method (ViT) captures long-range dependencies and drives the model to focus on diverse human-body parts. Our baseline fuses the local features extracted in the CNN structure and the long-range dependencies captured in the Transformer structure to enrich feature representations.

C. Experimental Results

We compare our approach with state-of-the-art VI-ReID methods on four datasets (Occluded-SYSU-MM01, Occluded-RegDB, SYSU-MM01 and RegDB). These competing methods are all based on the ResNet50 network, including expAT [55], HAT [56], AGW [10], DDAG [9], JSIA [57], Hi-CMD [58], CAJ [12], cm-SSFT [8] and CM-NAS [11]. These methods adopt the same experimental settings as in this paper.

Results on Occluded-SYSU-MM01: As shown in Table III, our approach is largely superior to the existing state-of-the-art methods in *all-search* and *indoor-search* evaluation modes. We observe that our approach achieves 41.49%/40.07 Rank-1/mAP in *all-search* mode, which outperforms the CM-NAS [11] method by 7.02%/5.22%. A similar phenomenon occurs with the *indoor-search* mode, our approach leads the CM-NAS method by 11.28% and 10.65% on Rank-1 and mAP.

Results on Occluded-RegDB: As shown in Table IV, our approach is compared with the state-of-the-art methods and significantly outperforms these methods in *visible-search-infrared* (V-I) and *infrared-search-visible* (I-V) evaluation modes. In V-I

TABLE II
CHARACTERISTICS OF OCCLUDED DATASETS

Dataset	Overall			Train			Test			Key Characteristic			
	IDs	VISs	NIRs	IDs	VISs	NIRs	IDs	Query	Gallery	Cams	Modes	Occluded Location	Occluded Area
Occluded-SYSU-MM01	491	30,071	15,792	395	22,258	11,909	96	3,803	301	6	All-search, Indoor-search	{t,b,l,r}	{1/2,1/4}
Occluded-RegDB	412	4,120	4,120	206	2,060	2,060	206	2,060	2,060	2	Visble-infrared, Infrared-visible	{t,b,l,r}	{1/2,1/4}

TABLE III
COMPARISON WITH STATE-OF-THE-ARTS ON THE OCCLUDED-SYSU-MM01 DATASET

Occluded-SYSU-MM01 Dataset									
Method	Source	All-Search				Indoor-Search			
		Rank-1	Rank-10	Rank-20	mAP	Rank-1	Rank-10	Rank-20	mAP
expAT [55]	TIP-2021	11.51	43.93	57.65	14.14	16.76	57.24	74.53	26.28
HAT [56]	TIFS-2020	12.46	43.68	61.24	14.30	14.49	55.57	75.95	24.29
AGW [10]	TPAMI-2021	18.20	53.46	69.71	20.88	22.24	67.57	82.61	32.27
DDAG [9]	ECCV-2020	20.56	59.58	74.44	23.56	24.37	68.48	83.24	35.14
JSIA [57]	AAAI-2020	22.76	65.47	79.34	23.41	25.26	72.48	86.76	35.63
Hi-CMD [58]	CVPR-2020	26.24	67.42	81.75	17.09	30.97	70.66	86.38	38.72
CAJ [12]	ICCV-2021	31.50	72.89	85.43	31.72	36.96	81.75	91.98	46.92
cm-SSFT [8]	CVPR-2020	34.33	74.20	82.71	33.56	38.83	78.21	88.18	40.83
CM-NAS [11]	ICCV-2021	34.47	75.18	86.67	34.85	34.01	78.03	90.26	43.34
Ours		**41.49**	**76.81**	**88.07**	**40.07**	**45.29**	**83.12**	**93.90**	**53.99**

TABLE IV
COMPARISON WITH THE STATE-OF-THE-ARTS ON THE OCCLUDED-REGDB DATASET. V-I MEANS VISIBLE-SEARCH-INFRARED MODE, WHILE I-V MEANS INFRARED-SEARCH-VISIBLE MODE

Occluded-RegDB Dataset									
Method	Source	V - I				I - V			
		Rank-1	Rank-10	Rank-20	mAP	Rank-1	Rank-10	Rank-20	mAP
HAT [56]	TIFS-2020	13.20	35.95	46.50	12.06	14.17	35.81	47.59	13.05
expAT [55]	TIP-2021	16.31	41.72	51.11	15.25	19.69	42.97	53.36	20.07
AGW [10]	TPAMI-2021	23.62	45.82	57.06	24.93	24.29	48.97	60.29	24.61
DDAG [9]	ECCV-2020	24.87	47.08	58.08	26.45	25.57	48.90	60.05	26.04
JSIA [57]	AAAI-2020	28.76	53.29	61.98	29.13	29.79	54.15	63.43	28.95
CM-NAS [11]	ICCV-2021	30.72	55.04	65.26	30.03	14.27	35.49	46.07	12.16
Hi-CMD [58]	CVPR-2020	31.74	53.63	65.94	32.81	32.68	55.74	68.65	31.69
CAJ [12]	ICCV-2021	36.64	62.37	73.17	35.49	36.01	64.82	74.91	33.81
cm-SSFT [8]	CVPR-2020	37.33	51.12	62.49	38.21	38.57	53.79	65.19	36.61
Ours		**52.14**	**72.33**	**82.57**	**52.44**	**51.84**	**69.13**	**80.69**	**50.94**

mode, our approach achieves 52.14% and 52.44% on Rank1 and mAP, outperforming the cm-SSFT [8] method by up to 14.81% and 14.23%. In I-V mode, our method leads the cm-SSFT method by 13.27% and 14.33% on Rank-1 and mAP.

Results on SYSU-MM01: The comparison results on SYSU-MM01 are shown in Table V. We can observe that the proposed approach achieves competitive performances compared with state-of-the-arts. Specifically, in the Rank-1/mAP, our approach achieves 72.98%/68.33% and 79.59%/82.95% in *all-search* and *indoor-search* modes, leads CAJ [12] method 3.10% and 3.33% in Rank-1 under two modes.

Results on RegDB: The comparison results on RegDB are shown in Table VI. The proposed approach outperforms existing SOTAs by large margins. Specifically, our approach achieves the Rank-1 accuracy of 88.44% and mAP of 87.37% in V-I mode, and Rank-1 accuracy of 87.03% and mAP of 86.20% in I-V mode, and significantly improving the Rank-1 by 3.41% and mAP 8.23% in V-I mode over the CAJ [12] method.

In addition, to our best knowledge, there are currently appeared two Transformer-based methods including SPOT [18] and DFLN-ViT [48] in the VI-ReID task. We have added a comparison of the two methods in Tables V and VI. As shown in Table V, we find that our approach outperforms DFLN-ViT and SPOT by a large margin in Rank-1 and mAP on the SYSU-MM01 dataset. As shown in Table VI, our approach leads SPOT method, and compared to DFLN-ViT, our method lags behind

TABLE V
COMPARISON WITH THE STATE-OF-THE-ARTS ON THE SYSU-MM01 DATASET

SYSU-MM01 Dataset					
Method	Source	All-Search		Indoor-Search	
		Rank-1	mAP	Rank-1	mAP
Hi-CMD [58]	CVPR-2020	34.94	35.94	-	-
JSIA [57]	AAAI-2020	38.10	36.90	43.80	52.90
expAT [55]	TIP-2021	38.57	38.61	-	-
AGW [10]	TPAMI-2021	47.50	47.65	54.17	62.97
DDAG [9]	ECCV-2020	54.75	53.02	61.02	67.98
HAT [56]	TIFS-2020	55.29	53.89	62.10	69.37
DFLN-ViT [48]	TMM-2022	59.84	57.70	62.13	69.03
cm-SSFT [8]	CVPR-2020	61.60	63.20	70.50	72.60
CM-NAS [11]	ICCV-2021	61.99	60.02	62.14	66.75
SPOT [18]	TIP-2022	65.34	62.25	69.42	74.63
CAJ [12]	ICCV-2021	69.88	66.89	76.26	80.37
Ours		**72.98**	**68.33**	**79.59**	**82.95**

TABLE VI
COMPARISON WITH THE STATE-OF-THE-ARTS ON THE REGDB DATASET

RegDB Dataset					
Method	Source	V - I		I - V	
		Rank-1	mAP	Rank-1	mAP
JSIA [57]	AAAI-2020	48.10	48.90	48.50	49.30
cm-SSFT [8]	CVPR-2020	61.60	63.20	70.50	72.60
expAT [55]	TIP-2021	67.45	66.51	66.48	67.31
DDAG [9]	ECCV-2020	69.34	63.46	68.06	61.80
Hi-CMD [58]	CVPR-2020	70.93	66.04	-	-
AGW [10]	TPAMI-2021	70.05	66.37	-	-
HAT [56]	TIFS-2020	71.83	67.56	70.02	66.30
SPOT [18]	TIP-2022	80.35	72.46	79.37	72.26
CM-NAS [11]	ICCV-2021	84.54	80.32	82.57	78.31
CAJ [12]	ICCV-2021	85.03	79.14	84.75	77.82
DFLN-ViT [48]	TMM-2022	**92.10**	82.11	**91.21**	81.62
Ours		88.44	**87.37**	87.03	**86.20**

TABLE VII
EVALUATION OF EACH MODULE OF OUR APPROACH ON THE OCCLUDED-SYSU-MM01 DATASET

Index	B	L	M	Occluded-SYSU-MM01 Dataset							
				All-search				Indoor-search			
				R1	R10	R20	mAP	R1	R10	R20	mAP
1	✓			34.08	66.53	77.39	32.53	40.72	76.54	86.32	46.40
2	✓	✓		39.23	71.79	83.30	38.66	43.47	80.07	90.72	51.79
3	✓		✓	38.52	72.31	81.62	37.37	42.66	79.35	90.13	50.67
4	✓	✓	✓	41.49	76.81	88.07	40.07	45.29	83.12	93.90	53.99

TABLE VIII
EVALUATION OF EACH MODULE OF OUR APPROACH ON THE OCCLUDED-REGDB DATASET

Index	B	L	M	Occluded-RegDB Dataset							
				V - I				I - V			
				R1	R10	R20	mAP	R1	R10	R20	mAP
1	✓			38.99	64.99	73.28	36.52	40.35	60.41	74.43	33.68
2	✓	✓		47.05	70.71	78.54	47.07	46.12	65.83	78.26	42.62
3	✓		✓	48.83	71.56	79.17	50.85	48.98	66.92	78.64	48.04
4	✓	✓	✓	52.14	72.33	82.57	52.44	51.84	69.13	80.69	50.94

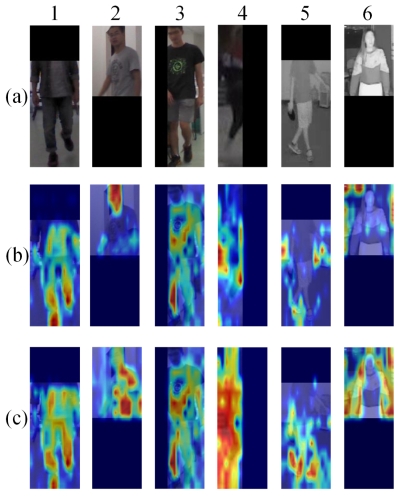


Fig. 7. The visualization of heatmaps on the Occluded-SYSU-MM01 dataset. (a) Input images, (b) Baseline, (c) Baseline + LFEM.

in Rank-1 but leads by a larger margin in mAP on the RegDB dataset.

The experimental results on the above four datasets (Occluded-SYSU-MM01, Occluded-RegDB, SYSU-MM01, and RegDB) show that our method outperforms on both occluded and non-occluded datasets, and existing methods perform poorly under different noise occlusions. The reason for this phenomenon is that the existing methods rely on the complete pedestrian image to construct a model. These methods are often disturbed by the noise in the occluded region when faced with occluded image, which degrades the recognition performance. Our approach achieves excellent performance by enhancing the saliency of local features and mining the relationship between local features to reduce the interference of noisy regions to improve feature discrimination.

D. Ablation Study

This subsection studies the effectiveness of each module involved in our approach on Occluded-SYSU-MM01 (*all-search* and *indoor-search modes*). "B" denotes the "Baseline" combining ResNet50 and ViT with the common learning objectives L_{id} and L_{tc}. "L" represents the local feature enhance module (LFEM), and "M" denotes the modality information fusion module (MIFM).

As shown in Tables VII and VIII, we observe that Baseline has already achieve much better performance than the compared models in Tables III and IV. The reason is that Baseline adopts CNN method (ResNet50) and Transformer method (ViT) as the basic structure, which incorporates the advantages of the CNN method (achieving some degree of shift, scale, and distortion invariance) and the advantages of the Transformer method (captures long-range dependencies and drives the model to attend to diverse human-body parts).

As shown in Table VII, the effectiveness of each component is revealed in *all-search* and *indoor-search*. Comparison index 1 with index 2, we observe that LFEM leads "B" by 5.15%/6.13% and 2.75%/5.39% in Rank-1/mAP in *all-search* and *indoor-search*. A similar increase also appears on the MIFM, the results of MIFM are 4.44%/4.84% and 1.94%/4.27% ahead of "Base" in Rank-1/mAP in *all-search* and *indoor-search*. Meanwhile, we observe a more significant improvement in the effect of using both LFEM and MIFM. In *all-search* mode on Rank-1 and mAP, an increase of 7.41% and 7.54%, respectively. As shown in Table VIII, a similar growth phenomenon also occurs in Occluded-RegDB dataset.

As shown in Fig. 7, we show the visualization diagram of the Baseline and Baseline + LFEM on the Occluded-SYSU-MM01 dataset. We observe that LFEM can effectively enhance the saliency of local features, e.g., for pedestrians 2, 4, 5 and 6, LFEM enhances local features such as shoulders, posture, legs and arms of pedestrians, respectively, and for pedestrians 1 and 3, LFEM significantly enhances local features including waist and chest of pedestrians.

Meanwhile, we notice that B+L tends to outperform B+M in the Occluded-SYSU-MM01 dataset, while B+M tends to outperform B+L in Occluded-RegDB dataset. We consider the following two reasons for this phenomenon: 1) as shown in Fig. 8,

TABLE IX
COMPARISON WITH STATE-OF-THE-ARTS ON THE OCCLUDED-SYSU-MM01 DATASET WITH DIFFERENT COLOR PATCHES

Occluded-SUYS-MM01 Dataset												
Occlusion type	Black patch				Gray patch				White patch			
Method	All-search		Indoor-search		All-search		Indoor-search		All-search		Indoor-search	
	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP
AGW [10]	18.20	20.88	22.24	32.27	19.22	22.18	24.09	34.26	18.96	22.28	23.01	32.71
CAJ [12]	31.50	31.72	36.96	46.92	33.00	34.82	37.55	44.38	30.50	31.90	37.46	45.79
CM-NAS [11]	34.47	34.85	34.01	43.34	38.61	40.52	42.19	50.78	30.64	32.88	37.76	44.37
Our	**41.49**	**40.07**	**45.29**	**53.99**	**42.02**	**43.11**	**45.15**	**55.46**	**39.65**	**40.53**	**42.26**	**50.76**

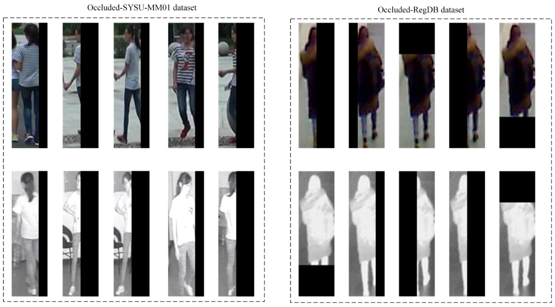


Fig. 8. The Example images on the Occluded-SYSU-MM01 and Occluded-RegDB dataset.

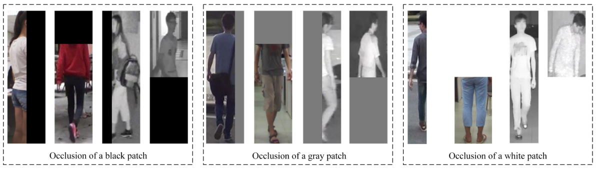

Occlusion of a black patch
Occlusion of a gray patch
Occlusion of a white patch

Fig. 9. Example of different color patches to occlude pedestrian images on the Occluded-SYSU-MM01 dataset.

we can observe that the variation of pedestrian's pose and actions is richer in Occluded-SYSU-MM01 dataset than in Occluded-RegDB dataset, and the sharpness of the images is significantly higher than in Occluded-RegDB dataset. LFEM needs to learn rich pedestrian features (containing the variation of pedestrian's actions and pose) to enhance the saliency representation of local features. 2) we observe that in the Occluded-RegDB dataset, the difference between the visible and NIR image modalities is much larger than the difference between the modalities in the Occluded-SYSU-MM01 dataset. The MIFM is more focused on handling the differences between modalities by co-modeling intra-modality information and inter-modality information interaction.

In summary, these meaningful improvements demonstrate that enhancing the saliency representation of local features and correlating the relationships between local features can enhance the discriminative power of global features. Intra-modality information enhancement and inter-modality information interaction effectively alleviate modality differences.

E. Discussion on Patches With Different Occlusion Colors

We validate the design of our approach by varying the occlusions color. As shown in Fig. 9, we use black, gray and white patches (random up and down, left and right positions, occlusion area size is random 1/2 and 1/4) to occlude pedestrian images on the SYSU-MM01 dataset. As shown in Table IX, the experimental results of our approach and the AGW [10], CAJ [12] and CM-NAS [11] methods under these three color occlusions. We observe that these methods perform better on gray occluded images than on the remaining two color occluded ones, while performing worse on white occluded images. Importantly, our method leads the other methods substantially on all three color occluded images. In our occluded VI-ReID, we use the occlusion of the black patch in our datasets.

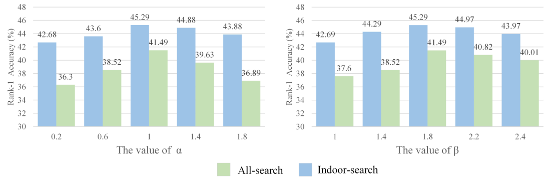


Fig. 10. Parameter analysis for α and β on the Occluded-SYSU-MM01 dataset.

F. Parameter Analysis

As shown in Fig. 10, we analyze the impact of different values of parameters α and β on the Occluded-SYSU-MM01 dataset. The parameters α and β control the weights of two modules in the total loss, which facilitate the learning of stronger discriminant features. By fixing the $\beta = 1.8$, we increase the parameter α from 0.2 to 1.8 to obtain the best setting α. Meanwhile, we fix the $\alpha = 1.0$ to change parameter β from 1.0 to 2.4. We can conclude that when $\alpha = 1.0$ and $\beta = 1.8$, *all-search* and *indoor-search* modes reach optimal performance with 41.49% and 45.29% in Rank-1.

G. Visualization

Visualization of Qualitative Results: As shown in Fig. 11, we demonstrate the Rank-10 retrieved results of selected several queries from the Occluded-SYSU-MM01 dataset. For each occluded query image on the left, the two rows of images on the right show top 10 matching images of our approach and "Baseline". We can observe that our approach effectively overcomes the noise interference in occluded regions and correctly identifies the same pedestrian image (green boxes). As a comparison, the "Baseline" is very sensitive to occlusion and returns a large number of falsely matched pedestrian images (red boxes). In summary, our approach is stable and robust on the Occluded-SYSU-MM01 dataset.

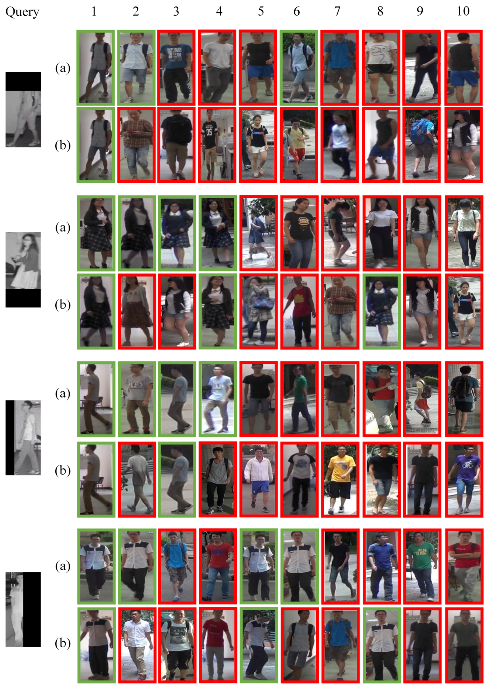


Fig. 11. The top ten retrieved results of four randomly selected images from the Occluded-SYSU-MM01 dataset (including quarter and half range occlusions in the upper, lower, left, and right regions of the image). (a) Our approach, (b) Baseline. The green boxes mean the correct matchings and the red boxes mean the wrong matchings. Each infrared query has four true matches at most.

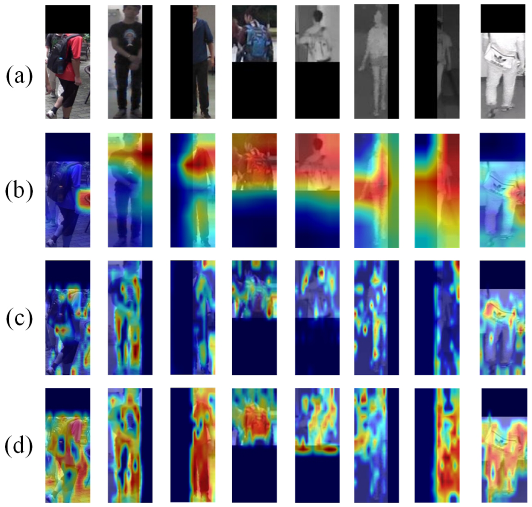


Fig. 12. The visualization of heatmaps on the Occluded-SYSU-MM01 dataset. (a) Input images, (b) ResNet50 method, (c) ViT method, (d) Our approach.

Visualization of Feature Heatmaps: We visualize features maps learned with Grad-CAN [59] by ResNet50 method based on CNN [43], ViT method based on Transformer [14] and our approach on the Occluded-SYSU-MM01 dataset. As shown in Fig. 12, we can find that the ResNet50 method is significantly disturbed by the occlusion noise, resulting in the inability to capture the effective local features in non-occluded regions. The Transformer method is able to capture local features with less noise interference, but cannot enhance local features in non-occluded regions. Our approach can effectively capture the features in non-occluded regions and enhance the saliency of the features, thus improving the feature discrimination ability.

VI. Conclusion

In this paper, we focus on realistic occluded visible-infrared person re-identification scenario, and introduce two occlusion datasets (Occluded-SYSU-MM01 and Occluded-RegDB) to simulate this scenario. Meanwhile, a matching framework has been proposed to tackle the Occluded VI-ReID in this scenario, which consists of local feature enhance module (LFEM) and modality information fusion module (MIFM). LFEM enhances the local feature saliency representation by exploring plentiful contextual information. MIFM constructs the dense intra-modality interaction mechanisms and the inter-modality interaction mechanism to mine intra-modality feature representations and fuse the information from two modalities, respectively.

Extensive experiments performed on four datasets (two occluded datasets and two non-occluded datasets) show that our approach achieves better performances than most state-of-the-art methods. The ablation studies sufficiently validated the effectiveness of each module of our approach. Meanwhile, in the section of visualization, it can be observed that our approach can focus more on the features of non-occluded regions and the performance of matching is excellent.

References

[1] P. Wang et al., "Deep multi-patch matching network for visible thermal person re-identification," *IEEE Trans. Multimedia*, vol. 23, pp. 1474–1488, 2021.

[2] Y. Huang et al., "Alleviating modality bias training for infrared-visible person re-identification," *IEEE Trans. Multimedia*, vol. 24, pp. 1570–1582, 2022.

[3] A. Wu, W.-S. Zheng, H.-X. Yu, S. Gong, and J. Lai, "RGB-infrared cross-modality person re-identification," in *Proc. Comput. Vis. Pattern Recognit.*, 2017, pp. 5390–5399.

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

[5] 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, 2018, pp. 7501–7508.

[6] M. Ye, Z. Wang, X. Lan, and P. C. Yuen, "Visible thermal person re-identification via dual-constrained top-ranking," in *Proc. Int. Joint Conf. Artif. Intell.*, 2018, vol. 1, pp. 1092–1099.

[7] 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.*, 2019, vol. 33, pp. 8385–8392.

[8] Y. Lu et al., "Cross-modality person re-identification with shared-specific feature transfer," in *Proc. Comput. Vis. Pattern Recognit.*, 2020, pp. 13376–13386.

[9] 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.*, 2020, pp. 229–247.

[10] M. Ye et al., "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.

[11] C. Fu et al., "CM-NAS: Cross-modality neural architecture search for visible-infrared person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 11823–11832.

[12] 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.*, 2021, pp. 13567–13576.

[13] A. Vaswani et al., "Attention is all you need," *Neural Inf. Process. Syst.*, vol. 30, pp. 6000–6010, 2017.

[14] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2020, pp. 1–6.

[15] N. Carion et al., "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.

[16] Q. Wu et al., "Discover cross-modality nuances for visible-infrared person re-identification," in *Proc. Comput. Vis. Pattern Recognit.*, 2021, pp. 4328–4337.

[17] 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, 2020.

[18] C. Chen et al., "Structure-aware positional transformer for visible-infrared person re-identification," *IEEE Trans. Image Process.*, vol. 31, pp. 2352–2364, 2022.

[19] 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.*, 2021, vol. 35, pp. 3520–3528.

[20] 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.*, 2021, pp. 12046–12055.

[21] H. Huang, D. Li, Z. Zhang, X. Chen, and K. Huang, "Adversarially occluded samples for person re-identification," in *Proc. IEEE/CVF Comput. Vis. Pattern Recognit.*, 2018, pp. 5098–5107.

[22] G. Wang et al., "High-order information matters: Learning relation and topology for occluded person re-identification," in *Proc. IEEE/CVF Comput. Vis. Pattern Recognit.*, 2020, pp. 6448-6457.

[23] J. Miao, Y. Wu, and Y. Yang, "Identifying visible parts via pose estimation for occluded person re-identification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 9, pp. 4624–4634, Sep. 2022.

[24] C. Yan et al., "Occluded person re-identification with single-scale global representations," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 11875–11884.

[25] M. Jia et al., "Matching on sets: Conquer occluded person re-identification without alignment," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, pp. 1673–1681.

[26] W.-S. Zheng et al., "Partial person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 4678–4686.

[27] L. He, J. Liang, H. Li, and Z. Sun, "Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, 2018, pp. 7073–7082.

[28] T. He, X. Shen, J. Huang, Z. Chen, and X.-S. Hua, "Partial person re-identification with part-part correspondence learning," in *Proc. IEEE/CVF Comput. Vis. Pattern Recognit.*, 2021, pp. 9101–9111.

[29] L. Gao et al., "Texture semantically aligned with visibility-aware for partial person re-identification," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 3771–3779.

[30] J. Zhuo, Z. Chen, J. Lai, and G. Wang, "Occluded person re-identification," in *Proc. Int. Conf. Multimedia Expo*, 2018, pp. 1–6.

[31] J. Miao, Y. Wu, P. Liu, Y. Ding, and Y. Yang, "Pose-guided feature alignment for occluded person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 542–551.

[32] M. Jia, X. Cheng, S. Lu, and J. Zhang, "Learning disentangled representation implicitly via transformer for occluded person re-identification," *IEEE Trans. Multimedia*, early access, Jan. 07, 2022, doi: 10.1109/TMM.2022.3141267.

[33] Y. Li et al., "Diverse part discovery: Occluded person re-identification with part-aware transformer," in *Proc. IEEE/CVF Comput. Vis. Pattern Recognit.*, 2021, pp. 2897–2906.

[34] J. Devlin, M.-W. Chang, K. Lee, and K. T., "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.

[35] H. Chen et al., "Pre-trained image processing transformer," in *Proc. IEEE/CVF Comput. Vis. Pattern Recognit.*, 2021, pp. 12294–12305.

[36] Y. Li, T. Yao, Y. Pan, and T. Mei, "Contextual transformer networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, 2022, doi: 10.1109/TPAMI.2022.3164083.

[37] S. Bhojanapalli et al., "Understanding robustness of transformers for image classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10231–10241.

[38] Z. Cao, C. Fu, J. Ye, B. Li, and Y. Li, "HiFT: Hierarchical feature transformer for aerial tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 15457–15466.

[39] Z. Sun, S. Cao, Y. Yang, and K. M. Kitani, "Rethinking transformer-based set prediction for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 3611–3620.

[40] R. Hu and A. Singh, "UniT: Multimodal multitask learning with a unified transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 1439–1449.

[41] H. Li, J. Xiao, M. Sun, E. G. Lim, and Y. Zhao, "Transformer-based language-person search with multiple region slicing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1624–1633, Mar. 2022.

[42] M. Sun, J. Xiao, E. G. Lim, S. Liu, and J. Y. Goulermas, "Discriminative triad matching and reconstruction for weakly referring expression grounding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 11, pp. 4189–4195, Nov. 2021.

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

[44] J. Deng et al., "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[45] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[46] Y. Wu and K. He, "Group normalization," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[47] X. He, Y. Zhou, Z. Zhou, S. Bai, and X. Bai, "Triplet-center loss for multi-view 3D object retrieval," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, 2018, pp. 1945–1954.

[48] J. Zhao et al., "Spatial-channel enhanced transformer for visible-infrared person re-identification," *IEEE Trans. Multimedia*, early access, Mar. 31, 2022, doi: 10.1109/TMM.2022.3163847.

[49] 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.

[50] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, pp. 13001–13008.

[51] J. Zhuo, J. Lai, and P. Chen, "A novel teacher-student learning framework for occluded person re-identification," 2019, doi: 10.48550/arXiv.1907.03253.

[52] C. Zhao et al., "Incremental generative occlusion adversarial suppression network for person REiD," *IEEE Trans. Image Process.*, vol. 30, pp. 4212–4224, 2021.

[53] P. Chen et al., "Occlude them all: Occlusion-aware attention network for occluded person re-id," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 11833–11842.

[54] D. Shen et al., "Es-Net: Erasing salient parts to learn more in re-identification," *IEEE Trans. Image Process.*, vol. 30, pp. 1676–1686, 2021.

[55] H. Ye, H. Liu, F. Meng, and X. Li, "Bi-directional exponential angular triplet loss for RGB-infrared person re-identification," *IEEE Trans. Image Process.*, vol. 30, pp. 1583–1595, 2021.

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

[57] G.-A. Wang et al., "Cross-modality paired-images generation for RGB-infrared person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 12144–12151.

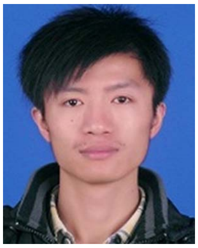
[58] S. Choi, S. Lee, Y. Kim, T. Kim, and C. Kim, "Hi-CMD: Hierarchical cross-modality disentanglement for visible-infrared person re-identification," in *Proc. IEEE/CVF Comput. Vis. Pattern Recognit.*, 2020, pp. 10254–10263.

[59] Jacob Gildenblat and contributors, "Pytorch library for cam methods," 2021. [Online]. Available: https://github.com/jacobgil/pytorch-grad-cam

Yujian Feng is currently working toward the Ph.D. degree with the School of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing, China. His research interests include pattern recognition, computer vision, machine learning, person re-identification, and machine learning.

Yimu Ji received the Ph.D. degree in computer science from the Nanjing University of Posts and Telecommunications (NJUPT), Nanjing, China, in 2006. He is currently a Professor with NJUPT. His research interests include intelligent driving, computer vision, and Big Data processing.

Fei Wu received the Ph.D. degree in information and communication engineering from the Nanjing University of Posts and Telecommunications (NJUPT), Nanjing, China, in 2016. He is currently with the College of Automation and Artificial Intelligence, NJUPT. His research interests include pattern recognition, artificial intelligence, and computer vision.



Guangwei Gao (Senior Member, IEEE) received the Ph.D. degree in pattern recognition and intelligence systems from the Nanjing University of Science and Technology, Nanjing, China, in 2014. He is currently with the College of Automation and Artificial Intelligence, Nanjing University of Posts and Telecommunications, Nanjing. His research interests include pattern recognition and computer vision.

Yang Gao is currently working toward the master's degree in computer science with the Nanjing University of Posts and Telecommunications, Nanjing, China. His research interests include pattern recognition and computer Vision.

Tianliang Liu received the Ph.D. degree in image processing and pattern recognition from Southeast University, Nanjing, China, in 2010. He is currently with the School of Communication and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing. His research interests include computer vision and pattern recognition.

Shangdong Liu is currently working toward the Ph.D. degree in computer system architecture with Southeast University, Nanjing, China. His research interests include computer vision, data analysis and processing, and cloud computing applications.

Xiao-Yuan Jing received the Doctoral degree in pattern recognition and intelligent system from the Nanjing University of Science and Technology, Nanjing, China, in 1998. He is currently a Professor with the School of Computer, Wuhan University, Wuhan, China, and Guangdong Provincial Key Laboratory of Petrochemical Equipment Fault Diagnosis, Guangdong University of Petrochemical Technology, Maoming, China. His research interests include pattern recognition and artificial intelligence.

Jiebo Luo (Fellow, IEEE) is currently a Professor of computer science with the University of Rochester, Rochester, NY, USA. He has authored more than 400 technical articles and holds more than 90 U.S. patents. His research interests include computer vision, NLP, machine learning, data mining, computational social science, and digital health. He has served on the Editorial Boards of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON BIG DATA, *ACM Transactions on Intelligent Systems and Technology*, *Pattern Recognition*, *Knowledge and Information Systems*, *Machine Vision and Applications*, and *Journal of Electronic Imaging*.