Intelligent Automation
& Soft Computing

Tech Science Press

DOI: 10.32604/iasc.2023.039894
Article

Pure Detail Feature Extraction Network for Visible-Infrared Re-Identification

Jiaao Cui[1], Sixian Chan[1,2,*], Pan Mu[1], Tinglong Tang[2] and Xiaolong Zhou[3]

[1]Zhejiang University of Technology, Hangzhou, 310023, China
[2]Hubei Key Laboratory of Intelligent Vision Based Monitoring for Hydroelectric Engineering, The College of Computer and Information, China Three Gorges University, Yichang, 443002, China
[3]The College of Electrical and Information Engineering, Quzhou University, Quzhou, 324000, China

*Corresponding Author: Sixian Chan. Email: sxchan@zjut.edu.cn

Received: 22 February 2023; Accepted: 15 May 2023; Published: 23 June 2023

Abstract: Cross-modality pedestrian re-identification has important applications in the field of surveillance. Due to variations in posture, camera perspective, and camera modality, some salient pedestrian features are difficult to provide effective retrieval cues. Therefore, it becomes a challenge to design an effective strategy to extract more discriminative pedestrian detail. Although many effective methods for detailed feature extraction are proposed, there are still some shortcomings in filtering background and modality noise. To further purify the features, a pure detail feature extraction network (PDFENet) is proposed for VI-ReID. PDFENet includes three modules, adaptive detail mask generation module (ADMG), inter-detail interaction module (IDI) and cross-modality cross-entropy (CMCE). ADMG and IDI use human joints and their semantic associations to suppress background noise in features. CMCE guides the model to ignore modality noise by generating modality-shared feature labels. Specifically, ADMG generates masks for pedestrian details based on pose estimation. Masks are used to suppress background information and enhance pedestrian detail information. Besides, IDI mines the semantic relations among details to further refine the features. Finally, CMCE cross-combines classifiers and features to generate modality-shared feature labels to guide model training. Extensive ablation experiments as well as visualization results have demonstrated the effectiveness of PDFENet in eliminating background and modality noise. In addition, comparison experiments in two publicly available datasets also show the competitiveness of our approach.

Keywords: Person re-identification; multimedia; image retrieval

1 Introduction

In recent years, extensive research [1,2] has begun to focus on the application of artificial intelligence in people's lives, such as healthcare, security, and transportation. Pedestrian retrieval is one of the critical problems in the security field, and its derived ReID task has essential research significance. VI-ReID aims to discover pedestrians with the same identity across different spectral cameras. Mining rich and modality-invariant pedestrian features are the key to solving this problem. Many early methods [3–5] directly pool the feature maps output by the network as pedestrian representations. However, this mode only preserves globally significant features, which loses some details (e.g., body characteristics, gender, clothing style) leading to a lack of discriminative features.

To preserve detailed information, Sun et al. [6] introduce a part pooling method, which extracts salient local features of pedestrians by limiting the pooling area. On this basis, Wang et al. [7] design three branches to perform part pooling with different steps to obtain multi-scale pedestrian features. Zhang et al. [8] propose an AlignReID strategy that computes the shortest path between part features of two images as a similarity metric. Part pooling alleviates the drawbacks of global pooling to a certain extent. When pedestrians are not covered with images, pooing is still disturbed by background information. Some methods [9–11] utilize pose estimation models to generate accurate human masks. However, these methods don't explicitly judge the reliability of joint predictions. Thus, these methods rely heavily on the accuracy of predictions.

Features, such as colorization, texture, and pedestrian appearance, change due to differences in the camera spectrum. This feature is called modality noise. VI-ReID requires features to contain as many modality-invariant features as possible. The current method takes the feature distance between modalities as the training target and designs the loss to suppress the modality-variant information in the features. However, as mentioned in literature [12], distance limitation is not conducive to the model learning sample distribution and affects the feature representation ability.

To address the above issues, a pure detail feature extraction network (PDFENet) is proposed. First, the adaptive detail mask generation module (ADMG) is introduced for solving the background noise problem introduced by part pooling. The ADMG adaptively generates the pedestrian detail mask based on the joint position derived from the pose estimation. Unlike the existing pose-based methods, ADMG explicitly evaluates the accuracy of each joint position, greatly alleviating the adverse effects caused by incorrect position. Due to the poor performance of the pose estimation model in the lower body, ADMG is only used to pool the upper body of the pedestrian, and part pooling is still utilized for the lower body. To purify the lower body features, the inter-detail interaction module (IDI) is further proposed, which utilizes the location information provided by ADMG to eliminate irrelevant backgrounds included in the part. IDI interacts with features' semantic and spatial information by modeling the correlation between detail features. Combined with ADMG and IDI, the background noise in pedestrian features is effectively suppressed. Finally, cross-modality cross-entropy (CMCE) loss is proposed to guide the model to filter modality noise from a representational learning perspective. Rather than indirectly limiting modality noise from feature distance, CMCE directly trains the model to filter the noise by generating modality-irrelevant pedestrian feature labels. By combining the above three modules, PDFENet can well purify pedestrian features. The main contributions of this paper are as follows:

1. The ADMG is proposed, which can adaptively generate masks of pedestrian details based on the features captured by the model. The mask is used to suppress background noise in the features.
2. The IDI is designed, which explicitly transmits semantic information according to detail relevance and further suppresses irrelevant background features.
3. The CMCE is established, which suppresses modality-variant noise by combining classifiers and features from different modalities. The classification results are used as soft labels to guide model training.

2 Related Work

2.1 Cross-Modality Pedestrian Re-ID

VI-ReID aims to match pedestrian images of different modalities captured by infrared and visible cameras. VI-ReID can be divided into three categories according to how modality differences are handled: pixel-level, feature-level, and a combination of both. Pixel-level methods are mainly based on GAN networks. Wang et al. [13] reduce modality differences by rendering visible images as infrared-style fake images through a GAN model. Wang et al. [14] entangle pixels according to modality-variant and modality-invariant and reconstructs the image with the help of the GAN network. Since the modality gap at the pixel level requires complex network structures for mapping, the quality of the generated images is difficult to guarantee.

At the feature-level, Zhang et al. [15] proposes to generate cross-modality fake feature vectors through GAN in the deep layers of the network. Compared with the shallow layers, the deep layers of the network contain more semantic information and the modality differences are reduced. Park et al. [5] propose to compute the dense correspondence of feature maps between modalities, and generate corresponding modality feature maps according to the relationship. Combining the two levels, Li et al. [16] first proposed to use a lighter-weight generative network to map visible images to an intermediate X-modality, and then guide the network to learn modality-invariant features from the feature-level. Zhang et al. [17] introduces two intermediate modalities to alleviate further the inter-modality differences at the pixel level of the image.

2.2 Part-Based Methods for Pedestrian Re-ID

Human detail features are considered to be important cues for establishing homogeneous pedestrian connections. To extract pedestrian detail features, Sun et al. [6] introduce the method of part pooling for the first time, which divides the feature map into strips in the vertical direction and pools them. In reality, pedestrians may only occupy a part of the image, so part pooling will inevitably introduce background information. To suppress background noise and enhance local features, some methods [18,19] introduce a pose estimation model, which helps the network to extract fine local features through its output heatmap. However, these methods ignore the prediction that may be wrong.

2.3 Loss Function

The loss function of VI-ReID is to train the network to learn pedestrian features with modality-invariant information. The current method designs the loss function from the perspective of classification and feature metrics. Some methods [20–23] treat pedestrian ReID as a classification problem and train the network with pedestrian IDs as labels. However, the label does not contain modality information, so it is difficult to train the network to learn modality-invariant features. Most methods [24–26] use the distance of features to measure the modality differences. Ye et al. [27] propose a BDTR loss to shorten the distance of the same pedestrians between modalities. Liu et al. [28] introduce a hetero-center triplet loss, which abandoned the traditional triplet and only limited the distance between modality feature centers. Ling et al. [29] propose the cross-modality earth mover distance (CM-EMD) loss, which suppresses the modality-variant information by reducing the optimal transport cost between features. Although these methods help the network alleviate modality differences to some extent, strict distance restrictions will prevent the model from perceiving the actual sample distribution.

3 Methodology

3.1 Overview

Backbone. As shown in Fig. 1, backbone of PDFENet includes a main branch and an auxiliary branch. The main branch contains a pre-trained ResNet-50 to extract the global features of pedestrians $F_g^m \in R^{\left(C\times H_g\times W_g\right)}(m \in \{V, I\})$. The auxiliary branch contains a pre-trained pose estimation model, which generates the heatmap $H^m \in R^{(16\times H\times W)}$ corresponding to the 16 pedestrian joints.

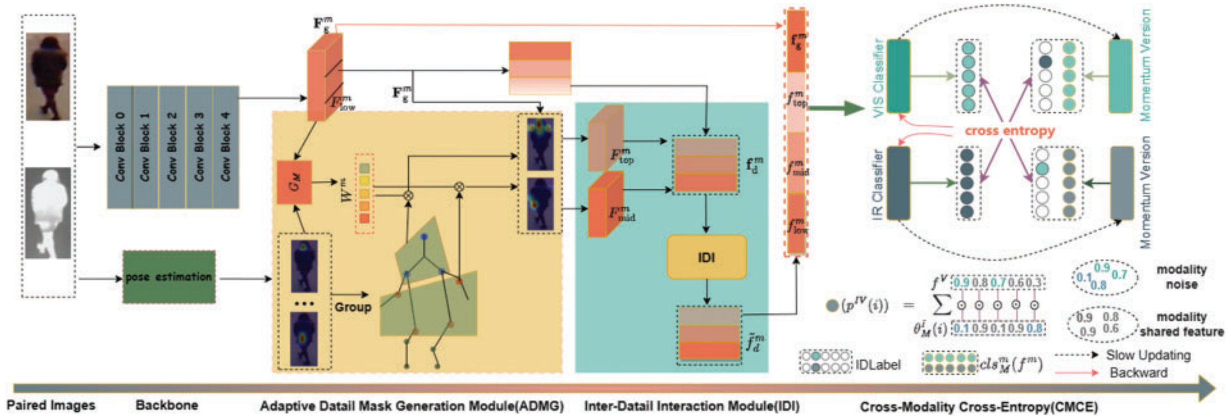

Figure 1: The structure of proposed PDFENet

Framework. After Backbone, ADMG, IDI, and CMCE are introduced to filter the background and modality noise in the features, respectively. First, ADMG learns the semantic correlation between H^m and F_g^m to evaluate the quality of each H^m and weights H^m to generate a mask of detailed features. The top and mid part of pedestrian features are filtered by mask and extracted by global pooling, and the lower part of pedestrian features are extracted by part pooing. Then, IDI is introduced to model the connections between detailed features and further suppress irrelevant background information. Finally, CMCE unites modality-specific classifiers and features to obtain modality-shared feature labels and guides the model to filter modality noise through the cross-entropy function. Beside CMCE, a center loss [30] function is employed to guide the model training.

3.2 Adaptive Detail Mask Generation Module (ADMG)

GCM [31] pretrained on MPII [32] is employed as our pose estimation model. Considering that GCM is affected by dataset and modality differences, random experiments is conducted on the SYSU-MM01 [33], and the results are shown in Table 1. The joint with more balanced accuracy between the two modalities is selected to participate in mask generation. The selected heatmap of the joint is denoted as $H^m \in R^{(9\times H\times W)}$. By spatial location, H^m are divided into two groups, $H_{top}^m \in R^{5\times H\times W}$ (thorax, upper-neck, head-top, left shoulder, right shoulder) and $H_{mid}^m \in R^{4\times H\times W}$ (left hip, left elbow, right elbow, left wrist). Since the pose estimation doesn't perform well on the lower body of pedestrians, part pooling is utilized extract the corresponding position features. Inspired by SENet [34], a G_M module is proposed, which contains a linear layer to sense the semantic connection between H^m and F_g^m. G_M perceives the semantic affiliation to determine the credibility of each heatmap. First, H^m are downsampled to the same size as F_g^m by max pooling. Connect H^m and F_g^m along the channel to obtain the pose feature map $F_{pose}^m \in R^{((C+9)\times H_g\times W_g)}$, and send F_{pose}^m to the G_M module. First, G_M performs global average pooling on F^m_{pose} to obtain the descriptor $GAP(F^m_{pose}) \in R^{((C+9)\times 1)}$ of each channel and captures the dependencies between channels through the linear layer:

$$W^m = Sigmoid\left(GAP\left(F^m_{pose}\right)^T W_1 + b_1\right) \tag{1}$$

where $W_1 \in R^{((C+9)\times 9)}$ and $b_1 \in R^{(1\times 9)}$ represent the parameters of the linear layer, and $W^m \in R^{(1\times 9)}$ represents the credibility of each heatmap in H^m. W^m is divided into $W^m_{top} \in R^{(1\times 5)}$, $W^m_{mid} \in R^{(1\times 4)}$. W^m are accumulated H^m to generate M^m_{top} and M^m_{mid} corresponding to each group. The formula is:

$$M^m_{top} = W^m_{top} H^m_{top} \tag{2}$$

$$M^m_{mid} = W^m_{mid} H^m_{mid} \tag{3}$$

where the heatmap is squeezed into two dimensions, $H^m_{top} \in R^{(5\times H_g W_g)}$, $H^m_{mid} \in R^{(4\times H_g W_g)}$. M^m_{top} and M^m_{mid} restores the spatial dimension $R^{(H_g\times W_g)}$. Finally, we enhance the detail information and suppress the background information through M^m_{top}, M^m_{mid}:

$$F^m_{\text{top}} = M^m_{\text{top}} \odot F^m_g \tag{4}$$

$$F^m_{\text{mid}} = M^m_{\text{mid}} \odot F^m_g \tag{5}$$

where $\odot$ represents the dot product operation. For the pedestrian lower body information, F^m_g are divided into three equal parts along the vertical direction, and the last part $F^m_{low} \in R^{(C\times \frac{H_g}{3}\times W_g)}$ is used as the lower body representation.

Table 1: The accuracy of the heatmap generated by GCM for 16 joints in two modalities

Joint	R-ankle	R-knee	R-hip	L-hip	L-knee	L-ankle	Pelvis	Thorax
VIS	70.3%	70.3%	43.7%	70.3%	65.6%	70.3%	43.7%	70.3%
IR	31.2%	31.2%	43.7%	65.6%	31.2%	31.2%	43.7%	65.6%
Joint	Upper-neck	Head-top	R-wrist	R-elbow	R-shoulder	L-shoulder	L-elbow	L-wrist
VIS	70.3%	65.6%	31.2%	65.6%	65.6%	70.3%	70.3%	65.6%
IR	65.6%	70.3%	43.7%	65.6%	70.3%	70.3%	65.6%	70.3%

3.3 Inter-Detail Interaction Module (IDI)

The global feature map F^m_g and the three detailed feature maps F^m_{top}, F^m_{mid}, and F^m_{low} are pooled to get their corresponding feature vectors $f^m_g \in R^{(C\times 1)}$, $f^m_{top} \in R^{(C\times 1)}$, $f^m_{mid} \in R^{(C\times 1)}$, and $f^m_{low} \in R^{(C\times 1)}$, respectively. As mentioned before, F^m_{low} contain background noise, while F^m_{top} and F^m_{mid} contain accurate pedestrian details. Therefore, IDI module is introduced, which exploits the affiliations between features to suppress background noise further. First, the three detail features are concatenated to get $f^m_d \in R^{(3\times C)}$. Then, three linear layers, $Q(\cdot)$, $K(\cdot)$, and $V(\cdot)$, are set up to get the query, key, and value corresponding to f^m_d. Query and key are multiplied to establish the semantic relationship between features. The value is accumulated according to the semantic relationship to obtain the purified feature $\hat{f}^m_d \in R^{3\times C}$.

The process is formulated as follows:

$$\hat{f}^m_d = \text{softmax}\left(\frac{Q\left(f^m_d\right)\left(K\left(f^m_d\right)\right)^T}{\sqrt{C}}\right) V\left(f^m_d\right) \tag{6}$$

Then, an adaptive weight is used to enhance the discriminative features. The weights are generated by two linear layers and a ReLU layer, and the calculation formula is:

$$W_d^m = \mathrm{ReLU}\left(\hat{f}_d^m W_2 + b_2\right) W_3 + b_3 \tag{7}$$

where $W_2 \in R^{(C\times\frac{C}{4})}$, $W_3 \in R^{(\frac{C}{4}\times 1)}$, $b_2 \in R^{(\frac{C}{4}\times 1)}$, $b_3 \in R^{(1\times 1)}$ represent linear layer parameters. We dot-multiply W_d^m with $\hat{f}_d^m$ to get the enhanced detail feature $\tilde{f}_d^m$.

$$\tilde{f}_d^m = W_d^m \odot \hat{f}_d^m \tag{8}$$

The global feature f_g^m and the detail feature $\tilde{f}_d^m \in R^{C\times 1}$ are concatenated along the channel as the final pedestrian representation $f^m \in R^{4C\times 1}$.

3.4 Cross-Modality Cross-Entropy (CMCE)

In classification tasks, cross-entropy measures how similar the predictions of the classifier are to the labels, guiding the model to learn identity features. We want the network to only focus on the identity features shared by the modalities, but the labels do not contain the feature information of the other modality. Therefore, we set up two modality-specific classifiers that only compute feature prediction results in a specific modality. The probability of the category predicted by the classifier is actually the cosine similarity of its parameters and features, as shown in Fig. 1. We use f^V, f^I, cls^V and cls^I to represent the pedestrian features and classifiers in the two modalities, respectively, and y to represent the one-hot identity label. Taking visible as an example, cls^I is utilized to predict the category f^V belongs to. The parameters of the cls^I reflect the typical identity characteristics of infrared modality, while f^V represents the pedestrian characteristics under visible. Therefore the cosine similarity between cls^I and f^V can suppress the modality-specific information to a certain extent. We denote $p^{m_1m_2} = Softmax(cls^{m_1}(f^{m_2})), m_1, m_2 \in \{V, I\}$. The cross-modality cross-entropy function is:

$$\mathcal{L}_M^V = -\sum\nolimits_{i=0}^{Np} p^{IV}(i)\log\left(p^{VV}(i)\right) \tag{9}$$

$$\mathcal{L}_M^I = -\sum\nolimits_{i=0}^{Np} p^{VI}(i)\log\left(p^{II}(i)\right) \tag{10}$$

where N_p is the number of pedestrian categories. Using p^{IV} and p^{VI} as the soft label can guide the model to discover features that coexist in two modalities. The identity cross-entropy also constrains f^V and f^I:

$$\mathcal{L}_{ID}^V = -\sum\nolimits_{i=0}^{Np} y(i)\log\left(p^{VV}(i)\right) \tag{11}$$

$$\mathcal{L}_{ID}^V = -\sum\nolimits_{i=0}^{Np} y(i)\log\left(p^{II}(i)\right) \tag{12}$$

To stably reflect the situation of feature learning under two modalities, the momentum versions of the modality classifier cls_M^V, cls_M^I are designed, whose parameters adopt a slow update strategy [35]:

$$\theta_M^V = \alpha\theta_M^V + (1-\alpha)\theta_C^V \tag{13}$$

$$\theta_M^I = \alpha\theta_M^I + (1-\alpha)\theta_C^I \tag{14}$$

where $\theta_C^{V/I}$ and $\theta_M^{V/I}$ are the parameters of $cls^{V/I}$ and $cls_M^{V/I}$, respectively, and α is the momentum coefficient. The momentum version of the classifier generates p^{IV} and p^{VI}. CMCE loss is summarized:

$$\mathcal{L}_{CM} = \lambda_1(\mathcal{L}_M^V + \mathcal{L}_M^I) + \lambda_2(\mathcal{L}_{ID}^V + \mathcal{L}_{ID}^I) \tag{15}$$

where λ_1 and λ_2 are hyperparameters to balance the contribution of each function.

3.5 Loss

Based on CMCE, a center loss is introduced to train the model, which relaxes the strict distance restriction between sample pairs with the help of feature centers. The feature center is calculated as follows:

$$C_p = \frac{1}{K} \sum\nolimits_{k=1}^{K} f_{p,k}^{m} \tag{16}$$

where C_p denotes the feature center of the p-th pedestrian, p ranges from 0 to N_p, and K denotes the number of samples labeled as $p. f_p^m$ denotes the feature of the pth pedestrian. The equation of the center contrast loss function is as follows:

$$\mathcal{L}_C = ||f_{p,k}^{m} - C_p||_2 + ||C_p - C_q||_2 \tag{17}$$

where C_p and C_q represent the feature centers of different pedestrians. Finally, overall loss function is:

$$\mathcal{L}_{total} = \mathcal{L}_{CM} + \mathcal{L}_C \tag{18}$$

4 Experiment

4.1 Datasets and Evaluation

Dataset: We evaluated PDFENet proposed by us on two mainstream public datasets, SYSU-MM01 [33] and RegDB [36]. SYSU-MM01 was the mainstream large Visible-Infraed database in VI-ReID tasks. The dataset contained 491 identities, including 29,033 RGB images and 15,712 IR images from 4 visible and 2 IR cameras in indoor and outdoor environments. The train set contained 395 identities, including 22,258 visible images and 11,909 IR images. The test set contains 96 identities, which were divided into query and gallery by modality. Among them, the query contained 3,803 IR images, and the gallery randomly selected 301 or 3010 images (single-shot or multi-shot) from visible images according to different test modalities. The gallery contained two camera selection modes: all-search (all cameras) and indoor-search (only indoor cameras).

The RegDB dataset consisted of 412 identities, each containing 10 Visible images and 10 Infrared images from a pair of overlapping VIS and IR cameras. We used the evaluation scheme in [37], which randomly divided the dataset into half identities for training and a half for testing. The experiment was repeated 10 times on the test set and the average was taken as the final test result.

Evaluation Protocol: Standard Cumulative Matching Characteristics (CMC) curve and mean Average Precision (mAP) were applied to evaluate the performance of our model.

4.2 Experimental Details

During the training phase, all images are randomly cropped to 384 × 128. The batch is set to 64 and contains four images for each modality of 8 pedestrians. The model is trained for a total of 240 epochs. The optimizer chooses Adam. The initial learning rate is set to 2×10^{-4}, which decays with a decay rate of 0.1 at 80 and 120 epochs, respectively. The weight decay is set to 5×10^{-4}. λ_1 and λ_2 are set to 2.5 and 0.5, respectively. Momentum coefficient α is set to 0.7. The backbone is ResNet-50, pre-trained on ImageNet. Baseline uses identity loss and center loss [30] on the backbone. In the test phase, we calculate the similarity between each query image and all images in the gallery and rank them. The rank result is used to calculate the evaluation protocol.

4.3 Comparison with the State-of-the-Art

Tables 2 and 3 show the experimental results. These methods are classified into three categories to mitigate the differences between modalities: pixel-level, feature-level, and a combination of both. For the pixel-level methods (JSIA [14], AlignGan [13]), the main idea is to generate pixel-level fake images to assist cross-modality retrieval. However, the sizeable modality difference at the pixel level requires a complex network structure for mapping, and fake images lose many pedestrian details. On SYSU-MM01, we achieved a 29.5\% lead on rank1 compared to AlignGan.

Table 2: Comparison of CMC (%) and mAP (%) performances with the state-of-the-art methods on SYSU-MM01

Method	Pub	All-search		Indoor-search	
		R1	mAP	R1	mAP
JSIA [14]	AAAI'20	38.10	36.90	43.80	52.90
AlignGAN [13]	ICCV'19	42.40	40.70	45.90	54.30
cm-SSFT(sq) [38]	CVPR'20	47.70	54.10	57.40	59.10
XIV-ReID [16]	AAAI'20	49.92	50.73	–	–
CMAlign [5]	ICCV'21	55.41	54.14	58.46	66.33
MSO [1]	MM'21	58.70	56.42	63.09	70.31
DG-VAE [39]	MM'20	59.49	58.46	–	–
MID [40]	AAAI'22	60.27	59.40	64.86	70.12
cm-SSFT [38]	CVPR'20	61.60	63.20	70.50	72.60
HCT [28]	TMM'20	61.68	57.51	63.41	68.17
SPOT [41]	TIP'22	63.34	62.25	69.42	74.63
MCLNet [4]	ICCV'21	65.40	61.98	72.56	76.58
FMCNet [15]	CVPR'22	66.34	62.51	68.15	74.09
Ours	–	**71.92**	**68.94**	**75.87**	**79.64**

Eliminating inter-modality differences at the feature level as the current mainstream approach achieves good results on VI-ReID. Method (MSO [1], HCT [28]) design the loss function to narrow the feature distance between different modalities. Method (CMAlign [5], FMCNet [15], DG-VAE [39]) learn modality-irrelevant features by establishing the mapping relationship of features between two modalities. These methods treat modality differences as feature distances to eliminate rather than guiding the model to learn modality-invariant features. Compared with the state-of-the-art method FMCNet, we can achieve a 5.5% Rank1 improvement and 6.43% mAP improvement. To combine the advantages of pixel and feature level, some methods (XIV-ReID [16], MID [40]) put forward the idea of X-modality. However, its essence still depends on feature distance to eliminate modality differences. Compared with the method MID, we achieve 11.6% and 9.5% improvement on Rank1 and mAP, respectively.

Table 3: Comparison of CMC (%) and mAP (%) performances with the state-of-the-art methods on RegDB

Method	Pub	Visible to infrared		Infrared to visible	
		R1	mAP	R1	mAP
JSIA [14]	AAAI'20	48.1	48.9	48.5	49.3
AlignGAN [13]	ICCV'19	57.9	53.6	56.3	53.4
CMM+CML [42]	MM'20	59.8	60.9	–	–
cm-SSFT(sq) [38]	CVPR'20	65.4	65.6	63.8	64.2
DG-VAE [39]	MM'20	73.0	71.8	–	–
MSO [1]	MM'21	73.6	66.9	74.6	67.5
CMAlign [5]	ICCV'21	74.17	67.64	72.43	65.46
SIM [43]	IJCAI'20	74.47	75.29	75.24	78.30
MCLNet [4]	ICCV'21	80.31	73.07	75.93	69.49
GECNet [44]	TCSVT'22	82.33	78.45	78.93	75.58
SPOT [41]	TIP'22	80.35	72.46	79.37	72.26
Ours	–	**83.06**	**80.07**	**82.57**	**79.79**

4.4 Ablation Experiments

In this section, extensive ablation experiments are set up to demonstrate the effectiveness of each module in our proposed PDFENet. In Table 4, we show experimental results on two datasets. On SYSU-MM01, we show the results in an all-search environment. The results of RegDB are displayed in the format of infrared to visible (visible to infrared).

Table 4: Ablation study in terms of CMC (%) and mAP (%) on SYSU-MM01 and RegDB

Components				SYSU-MM01		RegDB	
Baseline	ADMG	IDI	CMCE	R1	mAP	R1	mAP
√	×	×	×	56.77	55.96	68.88(67.33)	66.79(66.85)
√	√	×	×	63.31	61.89	73.35(73.30)	70.73(71.02)
√	×	×	√	63.54	59.61	75.83(73.45)	70.52(69.87)
√	√	×	√	68.08	65.44	78.40(79.56)	75.94(76.09)
√	√	√	√	**71.92**	**68.94**	**82.57(83.06)**	**79.79(80.07)**

4.4.1 Effectiveness of ADMG

As shown in Table 4, the detailed information extracted by ADMG is introduced based on the baseline, and the performance of the two datasets is significantly improved. To demonstrate the superiority of ADMG in dealing with background noise, it is compared with the mainstream part pooling methods, as shown in Table 5. The top, mid, and low correspond to the extraction methods of the three position details, respectively. For simplicity, we denote part pooling [6] as P, ADMG as

A. As shown in the red font in Table 5, as P is replaced with A in turn, the performance in the two datasets is also gradually improved. This proves that high-purity detail information can lead to better improvements than complete details with noise. Due to the poor performance of GCM in the lower body, the performance of P replaced A decreased. Therefore, we still use part pooling for the lower body. To visually demonstrate the effectiveness of ADMG, we visualize the detailed features enhanced by the $M^{m}_{top/mid}$ and the process of ADMG adaptive weighted $H^{m}_{top/mid}$, respectively, as shown in Fig. 2. As shown in Fig. 2b, for low-quality heatmap (red box), ADMG assigns a smaller weight to ensure the quality of the generated mask. This proves that ADMG can weaken the wrong localization by weight and preserve the accurate localization.

Table 5: The comparison between PCB and ADMG on SYSU-MM01 and RegDB

Components			SYSU-MM01		RegDB	
Top	Mid	Low	R1	mAP	R1	mAP
P	P	P	67.25	64.73	78.16(79.27)	75.07(76.25)
A	P	P	68.36	65.67	81.21(80.15)	78.27(76.90)
A	A	P	**71.92**	**68.94**	**82.57(83.06)**	**79.79(80.07)**
A	A	A	69.56	66.45	78.69(80.24)	76.01(76.59)

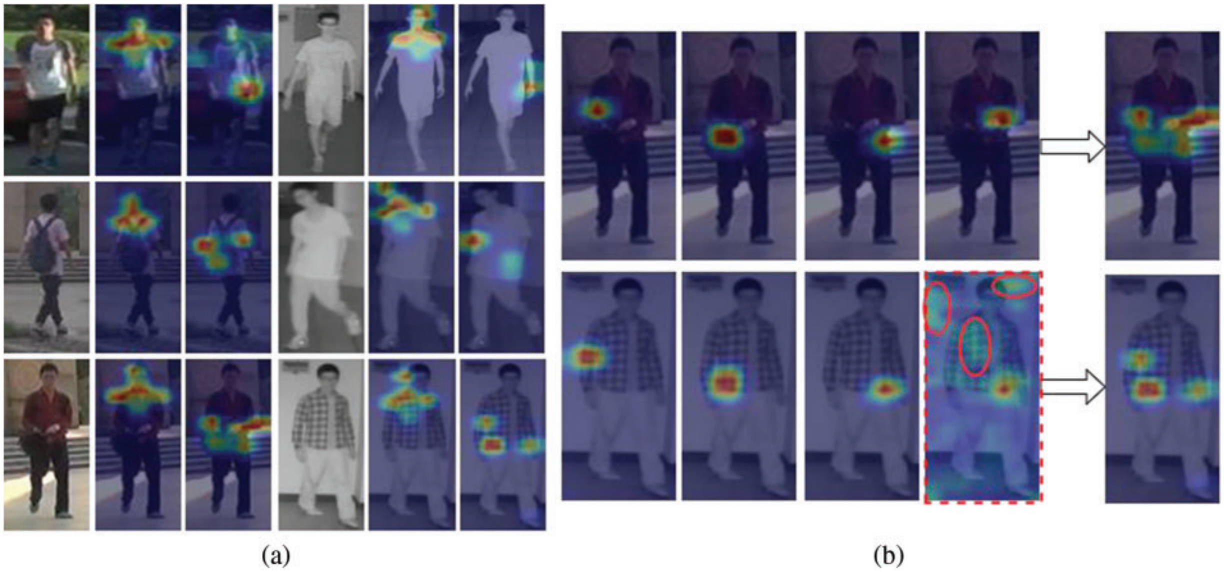

Figure 2: Visualization of ADMG on SYSU-MM01. (a) Detail features enhanced by mask. (b) Mask generated by ADMG adaptive weighted heatmap

4.4.2 Effectiveness of IDI

Table 4 shows that after introducing IDI, Rank1 increases by 3.84% and 4.17% (3.50%) on the two datasets, respectively. It is proved that IDI can further improve the representation ability of features by interacting the information between features.

4.4.3 *Effectiveness of CMCE*

CMCE are compared with the mainstream distance-based loss function. To be fair, we only replace the loss function based on the baseline. Table 6 shows our results. HCT abandons the strict distance limit of traditional triples and only limits the distance between modality centers, achieving good results. However, its center is based on the current batch calculation, which lacks representativeness. Compared with HCT, we achieved a 2.5% increase in Rank1 and a 1.5% increase in mAP. To intuitively feel the impact of distance limitation, we visualize the sample distribution of the network after different loss training. Fig. 3a shows the original feature distribution. Fig. 3b shows the distribution of samples learned by the network under the modality distance constraints. The distance between modalities within the same class is reduced, but the distance between classes is also narrowed. This proves that the distance limitation affects the model fitting the actual sample distribution. As shown in Fig. 3c, when CMCE is utilized, the intra-class sample points are evenly distributed, and the inter-class distances are well separated. Table 4 also shows that CMCE can significantly improve the performance of the model on both datasets.

Table 6: Comparison between CMCE and loss based on feature distance

Loss function	SYSU-MM01	
	R1	mAP
Baseline	56.77	55.96
B + expAT [45]	58.76	57.81
B + WRT [46]	59.03	57.81
B + HCT [28]	61.03	58.10
B + CMCE	**63.54**	**59.61**

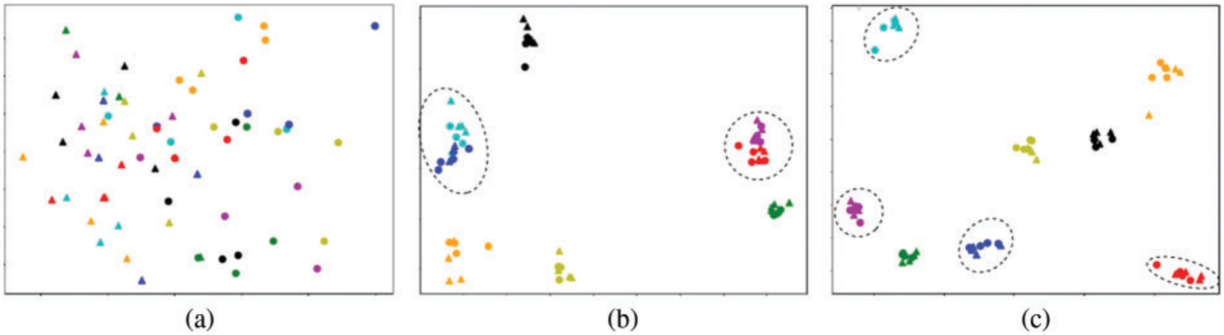

Figure 3: Visualization of features distribution, where colors represent pedestrian categories, circles represent visible images, and triangles represent infrared images. (a) Distribution of features perceived by the original model. (b) Distribution of features perceived by model after training based on modality distance. (c) Distribution of features perceived by model after CMCE training

4.4.4 *Visualization*

To further demonstrate the benefits of PDFENet, we have visualized the retrieval results of the Baseline model and PDFENet, respectively. The experiment is conducted in the multi-shot mode under the all-search environment of the SYSU-MM01. As shown in Fig. 4, each row represents the retrieval results of a pedestrian and is sorted from the largest to the smallest according to the similarity. Obviously, compared with the baseline, our model can find more target pedestrians and improve retrieval accuracy.

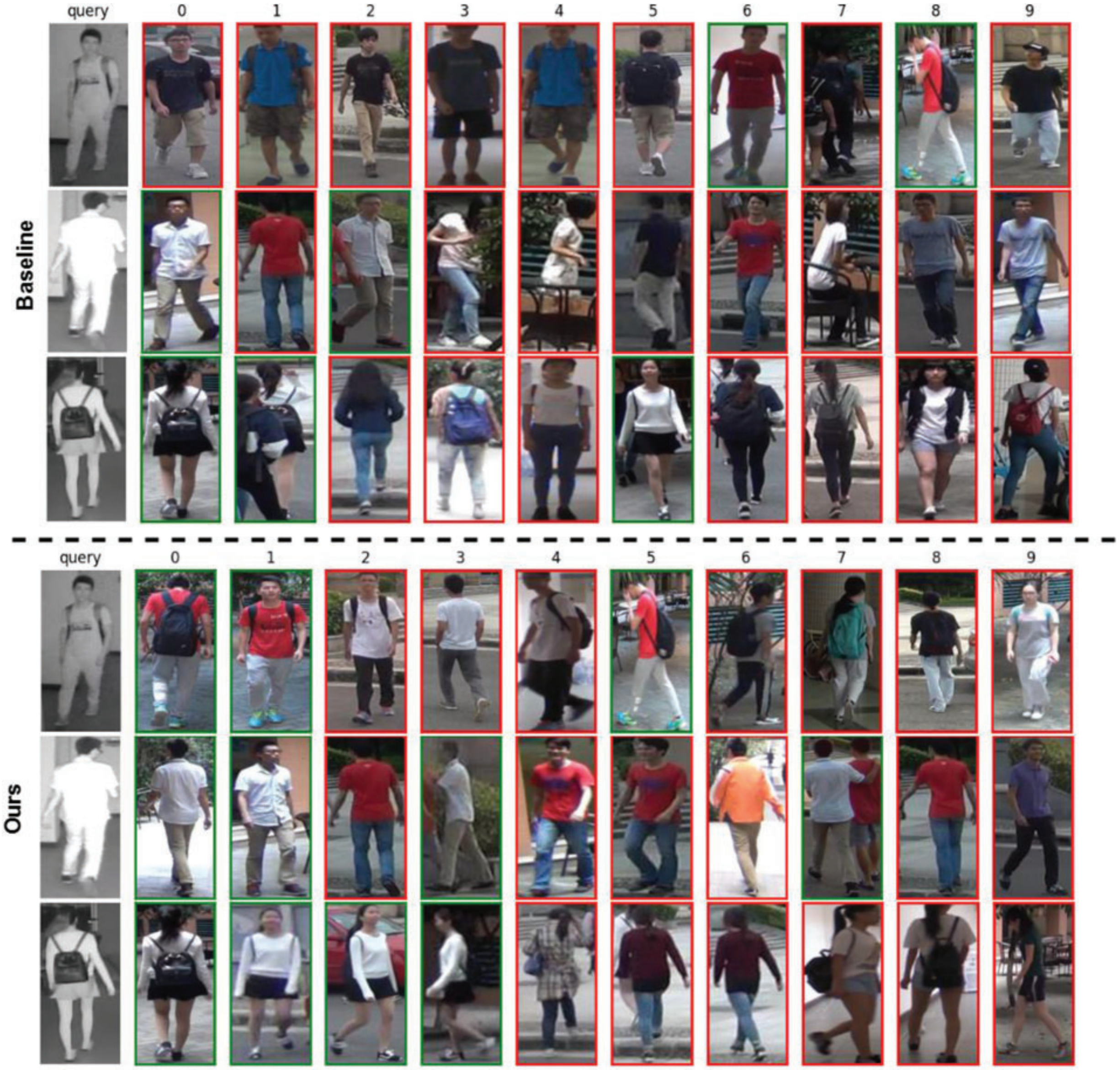

Figure 4: Compare the retrieval results of PDFENet and baseline on SYSU-MM01. Each row represents the retrieval result of a query image in the gallery. We show similar top 10 images in the gallery. The green box indicates successful retrieval, and the red box indicates failure

5 Citations

To suppress the background noise and modality noise contained in the feature, we proposed a pure detail feature extraction network (PDFENet). Against background noise, we utilized joint predictions generated by pose estimation to generate accurate detail masks. Mask can enhance corresponding details. Besides, we explicitly interacted with semantic information between details to suppress background noise further. For modality noise, we combined classifiers and features from different modalities and utilized the classification results as labels to guide network learning. Extensive experiments demonstrated the effectiveness of our method.

Funding Statement: This work is partially supported by the National Natural Science Foundation of China (Grant No. 61906168, 62202429), Zhejiang Provincial Natural Science Foundation of China (Grant No. LY23F020023), Construction of Hubei Provincial Key Laboratory for Intelligent Visual Monitoring of Hydropower Projects (2022SDSJ01).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] G. Dicuonzo, F. Donofrio, A. Fusco and M. Shini, "Healthcare system: Moving forward with artificial intelligence," *Technovation*, vol. 120, pp. 102510, 2023.

[2] A. A. Khan, A. A. Wagan, A. A. Laghari, A. R. Gilal, I. A. Aziz *et al.,* "Biomt: A state-of-the-art consortium serverless network architecture for healthcare system using blockchain smart contracts," *IEEE Access*, vol. 10, pp. 78887–78898, 2022.

[3] Y. Gao, T. Liang, Y. Jin, X. Gu, W. Liu *et al.,* "MSO: Multi-feature space joint optimization network for rgb-infrared person re-identification," in *Proc. MM*, Virtual Event, China, pp. 5257–5265, 2021.

[4] X. Hao, S. Zhao, M. Ye and J. Shen, "Cross-modality person re-identification via modality confusion and center aggregation," in *Proc. ICCV*, Montreal, QC, Canada, pp. 16383–16392, 2021.

[5] H. Park, S. Lee, J. Lee and B. Ham, "Learning by aligning: Visible-infrared person re-identification using cross-modal correspondences," in *Proc. ICCV*, Montreal, QC, Canada, pp. 12026–12035, 2021.

[6] 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. ECCV*, Munich, Germany, pp. 501–518, 2018.

[7] G. Wang, Y. Yuan, X. Chen, J. Li and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *Proc. MM*, Seoul, Republic of Korea, pp. 274–282, 2018.

[8] X. Zhang, H. Luo, X. Fan, W. Xiang, Y. Sun *et al.,* "Alignedreid: Surpassing human-level performance in person re-identification," *CoRR*, vol. abs/1711.08184, 2017.

[9] Z. He, H. Zhao and W. Feng, "Pgganet: Pose guided graph attention network for person reidentification," *CoRR*, vol. abs/2111.14411, 2021.

[10] Z. Ma, Y. Zhao and J. Li, "Pose-guided inter- and intra-part relational transformer for occluded person re-identification," in *Proc. MM*, Virtual Event, China, pp. 1487–1496, 2021.

[11] Y. Suh, J. Wang, S. Tang, T. Mei and K. M. Lee, "Part-aligned bilinear representations for person re-identification," in *Proc. ECCV*, Munich, Germany, pp. 418–437, 2018.

[12] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, M. Xu *et al.,* "Dual-path convolutional image-text embeddings with instance loss," *ACM Trans. Multim. Comput. Commun. Appl.*, vol. 16, no. 2, pp. 51:1–51:23, 2020.

[13] G. Wang, T. Zhang, J. Cheng, S. Liu, Y. Yang *et al.,* "Rgb-infrared cross-modality person reidentification via joint pixel and feature alignment," in *Proc. ICCV*, Seoul, Korea (South), pp. 3622–3631, 2019.

[14] G. Wang, T. Zhang, Y. Yang, J. Cheng, J. Cheng *et al.,* "Cross-modality paired-images generation for rgb-infrared person re-identification," in *Proc. AAAI*, New York, NY, USA, pp. 12144–12151, 2020.

[15] Q. Zhang, C. Lai, J. Liu, N. Huang and J. Han, "Fmcnet: Feature-level modality compensation for visible-infrared person re-identification," in *Proc. CVPR*, New Orleans, LA, USA, pp. 7339–7348, 2022.

[16] D. Li, X. Wei, X. Hong and Y. Gong, "Infrared-visible cross-modal person re-identification with an X modality," in *Proc. AAAI*, New York, NY, USA, pp. 4610–4617, 2020.

[17] Y. Zhang, Y. Yan, Y. Lu and H. Wang, "Towards a unified middle modality learning for visible-infrared person re-identification," in *Proc. MM*, Virtual Event, China, pp. 788–796, 2021.

[18] T. Wang, H. Liu, P. Song, T. Guo and W. Shi, "Pose-guided feature disentangling for occluded person re identification based on transformer," in *Proc. AAAI*, Virtual Event, pp. 2540–2549, 2022.

[19] Y. Miao, N. Huang, X. Ma, Q. Zhang and J. Han, "On exploring pose estimation as an auxiliary learning task for visible-infrared person re-identification," *CoRR*, vol. abs/2201.03859, 2022.

[20] L. Zheng, Y. Yang and A. G. Hauptmann, "Person re-identification: Past, present and future," *CoRR*, vol. abs/1610.02984, 2016.

[21] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang *et al.,* "Person re-identification in the wild," in *Proc. CVPR*, Honolulu, HI, USA, pp. 3346–3355, 2017.

[22] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, Z. Hu *et al.,* "Improving person re-identification by attribute and identity learning," *Pattern Recognit.*, vol. 95, pp. 151–161, 2019.

[23] T. Matsukawa and E. Suzuki, "Person re-identification using CNN features learned from combination of attributes," in *Proc. ICPR*, Cancún, Mexico, pp. 2428–2433, 2016.

[24] R. Hadsell, S. Chopra and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. CVPR*, New York, NY, USA, pp. 1735–1742, 2006.

[25] S. Chopra, R. Hadsell and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. CVPR*, San Diego, CA, USA, pp. 539–546, 2005.

[26] K. Q. Weinberger, J. Blitzer and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," in *Proc. NIPS*, Vancouver, British Columbia, Canada, pp. 1473–1480, 2005.

[27] M. Ye, Z. Wang, X. Lan and P. C. Yuen, "Visible thermal person re-identification via dual-constrained top-ranking," in *Proc. IJCAI*, Stockholm, Sweden, pp. 1092–1099, 2018.

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

[29] Y. Ling, Z. Zhong, D. Cao, Z. Luo, Y. Lin *et al.,* "Cross-modality earth mover's distance for visible thermal person re-identification," *CoRR*, vol. abs/2203.01675, 2022.

[30] Y. Wen, K. Zhang, Z. Li and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. ECCV*, Amsterdam, The Netherlands, vol. 9911, pp. 499–515, 2016.

[31] D. Osokin, "Global context for convolutional pose machines," *CoRR*, vol. abs/1906.04104, 2019.

[32] M. Andriluka, L. Pishchulin, P. V. Gehler and B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *Proc. CVPR*, Columbus, OH, USA, pp. 3686–3693, 2014.

[33] A. Wu, W. 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.

[34] J. Hu, L. Shen and G. Sun, "Squeeze-and-excitation networks," in *Proc. CVPR*, Salt Lake City, UT, USA, pp. 7132–7141, 2018.

[35] K. He, H. Fan, Y. Wu, S. Xie and R. B. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. CVPR*, Seattle, WA, USA, pp. 9726–9735, 2020.

[36] D. T. Nguyen, H. G. Hong, K. 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, pp. 605, 2017.

[37] M. Ye, X. Lan, J. Li and P. C. Yuen, "Hierarchical discriminative learning for visible thermal person reidentification," in *Proc. AAAI*, New Orleans, Louisiana, USA, pp. 7501–7508, 2018.

[38] Y. Lu, Y. Wu, B. Liu, T. Zhang, B. Yi *et al.,* "Cross-modality person re-identification with shared-specific feature transfer," in *Proc. CVPR*, Seattle, WA, USA, pp. 13376–13386, 2020.

[39] N. Pu, W. Chen, Y. Liu, E. M. Bakker and M. S. Lew, "Dual gaussian-based variational subspace disentanglement for visible-infrared person re-identification," in *Proc. MM*, Virtual Event/Seattle, WA, USA, pp. 2149–2158, 2020.

[40] Z. Huang, J. Liu, L. Li, K. Zheng and Z. Zha, "Modality-adaptive mixup and invariant decomposition for rgb-infrared person re-identification," in *Proc. AAAI*, Virtual Event, pp. 1034–1042, 2022.

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

[42] Y. Ling, Z. Zhong, Z. Luo, P. Rota, S. Li *et al.,* "Class-aware modality mix and center-guided metric learning for visible-thermal person re-identification," in *Proc. MM*, Virtual Event/Seattle, WA, USA, pp. 889–897, 2020.

[43] M. Jia, Y. Zhai, S. Lu, S. Ma and J. Zhang, "A similarity inference metric for rgb-infrared crossmodality person re-identification," in *Proc. IJCAI*, Yokohama, JPN, pp. 1026–1032, 2020.

[44] X. Zhong, T. Lu, W. Huang, M. Ye, X. Jia *et al.,* "Grayscale enhancement colorization network for visible-infrared person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1418–1430, 2022.

[45] 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.

[46] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao *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, 2022.