Information Fusion 103 (2024) 102128

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Multi-modal person re-identification based on transformer relational regularization

Xiangtian Zheng [a,b,*], Xiaohua Huang [a], Chen Ji [b], Xiaolin Yang [c], Pengcheng Sha [d,e], Liang Cheng [b]

[a] School of Computer Engineering, Nanjing Institute of Technology, Nanjing, 211167, China
[b] School of Geography and Marine Science, Nanjing University, Nanjing, 210023, China
[c] China Academy of Safety Science and Technology, Beijing, 100012, China
[d] School of Earth Science and Engineering, Hohai University, Nanjing, 211100, China
[e] GFZ German Research Centre for Geoscience, Department of Geodesy, Potsdam, 14473, Germany

ARTICLE INFO

Keywords:
Infrared pedestrian re-identification
Feature fusion
Multimodal person re-identification
Feature interaction

ABSTRACT

For robust multi-modal person re-identification (re-ID) models, it is crucial to effectively utilize the complementary information and constraint relationships among different modalities. However, current multi-modal methods often overlook the correlation between modalities at the feature fusion stage. To address this issue, we propose a novel multimodal person re-ID method called Transformer Relation Regularization (TRR). Firstly, we introduce an adaptive collaborative matching module that facilitates the exchange of useful information by mining feature correspondences between modalities. This module allows for the integration of complementary information, enhancing the re-ID performance. Secondly, we propose an enhanced embedded module that corrects general information using discriminative information within each modality. By leveraging this approach, we improve the model's stability in challenging multi-modal environments. Lastly, we propose an adaptive triple loss to enhance sample utilization efficiency and mitigate the problem of inconsistent representation among multimodal samples. This loss function optimizes the model's ability to distinguish between different individuals, leading to improved re-ID accuracy. Experimental results on several challenging visible-infrared person re-ID benchmark datasets demonstrate that our proposed TRR method achieves optimal performance. Additionally, extensive ablation studies validate the effective contribution of each component to the overall model. In summary, our proposed TRR method effectively leverages complementary information, addresses the correlation between modalities, and improves the re-ID performance in multi-modal scenarios. The results obtained from various benchmark datasets and the comprehensive analysis support the efficacy of our approach.

1. Introduction

Person re-ID aims to match the same individual across diverse cameras and time, making it applicable in real-world scenarios like video surveillance and smart security systems. With the rapid advancements in deep learning, person re-ID has achieved remarkable success in terms of accuracy and practical implementation. However, the majority of current re-ID methods are designed for visible environments, specifically during daytime or under well-lit conditions. Consequently, these methods often fail to operate effectively in low-light or nighttime settings. To address this challenge, incorporating infrared camera imagery from real-world scenes becomes crucial in facilitating re-ID across different modalities, leading to the emergence of Visible-Infrared Person Re-Identification (VI-ReID).

As shown in Fig. 1, occlusion, scale variation, and alignment problems are present in visible person re-ID, which are also faced by VI-ReID. In addition, VI-ReID suffers from modal mismatch and interference from feature fusion strategies. all these factors make VI-ReID more challenging. Researchers have proposed a number of approaches to address the above challenges. They mainly focus on solving the VI-ReID intra-modal and inter-modal feature variation problems. Some methods [1–4] use a two-stream structure with no parameter sharing to extract different modal features, and they adapt the model to modal differences by constraining the distributional similarity of the outputs. Such methods are good at extracting intra-modal features, but underutilise shared and complementary features between modes. Some of the recent approaches [5–7] try to segment the person image fixedly into multiple horizontal stripes or to guide the model to extract salient

* Corresponding author at: School of Computer Engineering, Nanjing Institute of Technology, Nanjing, 211167, China.
E-mail address: zxt@njit.edu.cn (X. Zheng).

https://doi.org/10.1016/j.inffus.2023.102128
Received 18 September 2023; Received in revised form 25 October 2023; Accepted 3 November 2023
Available online 4 November 2023

Fig. 1. In the real world, (a) and (b) pedestrian feature interference due to occlusion, scale and pose changes in a visible light environment. (c) and (d) in low-light environments still suffer from noise interference, pose alignment, and missing features. Compared to visible light person re-ID, VI-ReID is more difficult.

features with the help of attentional mechanisms. By utilising this information, the model's ability to perceive and adapt to feature differences between modalities can be enhanced. However, direct segmentation methods may introduce segmentation errors that can lead to the loss or incorrect representation of some body features. Attention-based methods, on the other hand, are highly sensitive to noise and disturbances in the image [8,9]. For example, occlusions or image quality problems can seriously affect the effectiveness of attention [10,11].

We propose a multi-modal person re-identification (re-ID) method based on Transformer Relation Regularization (TRR) to address the aforementioned problem. As depicted in Fig. 2, we introduce a greyscale base module to mitigate the inconsistency between visible and infrared samples. This module utilizes greyscale images as auxiliary samples to emphasize pedestrian contours and structural features. Furthermore, we propose a collaborative matching module to adaptively leverage the complementary information between modalities. This module captures unique information and clues pertaining to the characteristics of different modalities, compensating for their respective limitations. Additionally, we propose an enhanced embedding module to improve the feature's resistance to interference. This is achieved by refining integrated features with intra-modal discriminative features. Moreover, we incorporate global and local information into the final feature representation to enhance its overall representation. Finally, we address the issue of unreliable samples by proposing an adaptive triplet loss that independently leverages sample relationships. Through extensive experiments conducted on the SYSU-MM01 [12] and RegDB [13] datasets, our model achieves promising results and outperforms the majority of state-of-the-art methods. In summary, the main contributions of this paper can be summarized as follows:

(1) A collaborative matching module is designed to enhance the information of the modality by adaptively mining the complementary information between features. It compensates the limitations of each modality and improves stability of models in complex environments.

(2) We designed the enhanced embedding module. The interaction of discriminative and integrated information is utilised to suppress the pervasive noise problem. The completeness of features within a modality is improved by the fusion of global and local features.

(3) An adaptive triplet loss is proposed to enable the model to adaptively adjust the sample difficulty based on the sample relationship.

(4) The proposed TRR achieves optimal performance results on several challenging VI-ReID benchmark datasets. Extensive ablation studies validate the effective contribution of each component to the model.

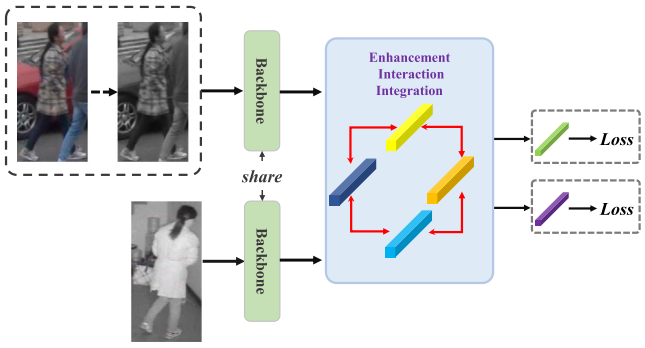

Fig. 2. Proposed multi-modal person re-ID method based on transformer relation regularization. Adaptive enhancement, interaction and integration of multi features between and within modalities improves the robustness of the model.

2. Related work

2.1. Visible-infrared person re-ID

The significance of capturing scenes is not limited to daylight alone; it extends to nighttime as well, forming an equally important aspect of a person's life. In situations with insufficient natural light, the use of infrared cameras becomes essential to capture images. In the context of VI-ReID, researchers have introduced numerous methods.

AGM [14] presents an innovative approach that employs a unified intermediate image space for the integration of multimodal information. This transformation converts the heterogeneous modal learning problem into a homogeneous grayscale modal learning problem, resulting in a significant improvement in feature representation by reducing modal differences within the image space. This approach closely aligns with the core principles of our methodology. CM-EMD [15] places its primary focus on the minimization of disparities between cross-modal features by assigning greater importance to feature pairs with lower internal variation. To attain more discriminative feature representations, CM-EMD [15] effectively reduces the ratio between variance within the same identity and variance across different identities. Moreover, CM-EMD introduces a multi-granularity structure, facilitating modal alignment at both coarse and fine-grained levels, which harmonizes with our objective of enhancing multimodal alignment. CycleTrans [16] extracts rich semantic information from features characterized by high modal correlation, utilizing pseudo-queries. Subsequently, these features undergo transformation into neutral features through modality-independent prototypes [17]. Moreover, CycleTrans [16] incorporates a feature cycle construction to enhance feature recognizability, aligning seamlessly with our goal of improving feature recognizability. KDEM [18] focuses on harnessing attribute knowledge correspondence across modalities. It accomplishes this by creating an augmented modality through the extraction and fusion of key semantic patterns from cross-modal representations, enriching feature expressiveness. KDEM also employs diverse loss functions to address redundancy in knowledge and strives to enhance feature semanticity and diversity, which is in line with our objective of enhancing feature expression and diversity. MBCE [19] is an unsupervised approach that amalgamates modality awareness and clustering centers to enhance the quality of clustering comparison and mutual information learning. MBCE introduces a cross-modal alignment method that leverages historical and recently acquired clustering agents to suppress inter-modal differences, ultimately improving inter-modal alignment. MRCN [20] focuses on addressing disparities between cross-modal features in VI-ReID through modal repair and compensation. Specifically, MRCN [20] restores normalized features by incorporating modality-independent features to recover missing information between modalities. Concurrently, modality-related features are employed to enhance inter-modal

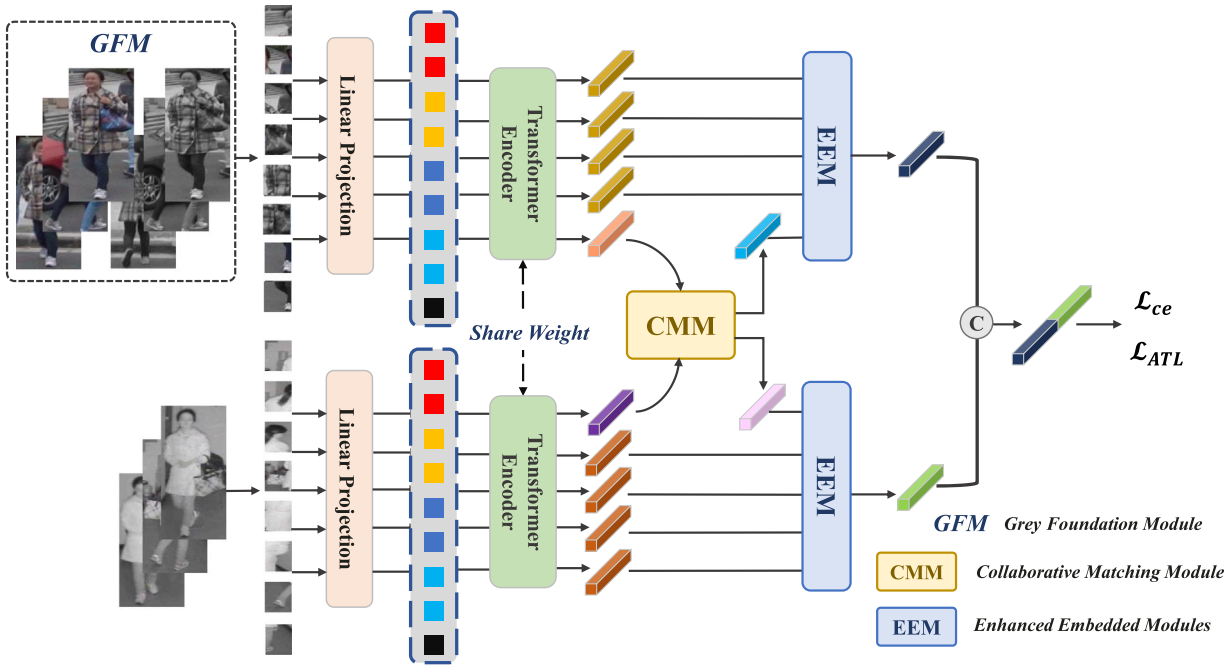

Fig. 3. Structural diagram of the TRR. It is a two-branch structure consisting of three modules. The greyscale base module takes the greyscale image along with the infrared image for network base training. The greyscale image is then replaced with a visible image for secondary training with the infrared image. The collaborative matching module and enhancement embedding module correspond to inter-modal and intra-modal feature interaction and enhancement, respectively. Finally, the output features are concatenated to form the final features.

consistency. To facilitate a better understanding of the relationship between these features, MRCN introduces a central quadratic causal loss [21], enhancing the capability to efficiently distinguish and utilize these features. PGM [22] reimagines the correlation between cross-modalities as a graph matching problem and introduces a matching cost to gauge cluster dissimilarity. By minimizing this matching cost, PGM [22] enhances the quality of global information. This strategic approach closely aligns with the focus of our research.

2.2. A visual transformer-based person re-ID

The constraints of the receptive field pose limitations on the performance of traditional CNN-based methods for person re-identification, particularly in capturing global dependencies. In contrast, the Visual Transformer (Vit) approach [23] exhibits significant promise in addressing long-range dependencies and elevating feature representation. ViT has already demonstrated remarkable achievements in computer vision tasks, including image classification, semantic segmentation, and video understanding.

PFD [24] employs a sliding window technique to divide the image into smaller chunks, capturing contextual relationships between modules using the Visual Transformer (ViT). Simultaneously, a pose-guided matching and distribution mechanism leverages the pose heatmap within the decoder as the key and value. This allows PFD to acquire a set of semantic part views, thereby enhancing the differentiation of visible body parts. To address the challenge of person occlusion, DPM [25] utilizes hierarchical semantic information for selecting visible parts from overall prototype and occlusion view feature representations. Additionally, DPM introduces a head enrichment module based on normalization and orthogonality constraints. This module not only enhances feature representation capabilities but also improves the model's capacity to suppress noise. In order to overcome the limitations of Transformers in extracting local features, PFT [26] introduces a learnable enhancement patch. Moreover, through feature slicing, fusion, and splicing, PFT ensures that the model not only learns long-range correlations between regions but also prioritizes attention to local features. PADE [27] extends data representation by incorporating cropped and erased versions of the original images. These processed images, along with the original image, are then input into a ViT-based multi-branch parameter-sharing network for feature extraction. Within this framework, the branches processing raw images enhance both global and local feature representations, akin to multimodal interaction. Ultimately, these features are concatenated to create the final character description. Pirt [28] introduces a two-branch pose estimation model. In this model, ResNet-50 [29] guides the extraction of internal features [30], while ViT is employed to construct relational features between body parts, enabling long-term part perception. DRL-Net [31] generates augmented samples with random obstacles and incorporates a semantic feature extraction layer based on a combination of CNN and Transformer architectures. This addition enhances the model's resilience to noise through the comparison of positive and negative samples.

3. Methodology

3.1. Overview

The overall structure of TRR is shown in Fig. 3, which is a two-branch structure consisting of a grey foundation module, a collaborative matching module, and an enhanced embedded module. The feature extractor is weight sharing Vit for capturing the base pedestrian features. The greyscale base module uses the greyscale image as a transition modality. The collaborative matching module compensates the limitations of each modality, and the enhanced embedded module enhances the completeness and richness of the features and improves the ability to suppress noise. Adaptive triplet loss solves the problem of unreliable samples.

3.2. Grey foundation module

Parameter sharing structures are commonly employed due to their effectiveness in learning shared features. However, they have the drawback of being less proficient in capturing modal-specific characteristics and may introduce redundant information. Grayscale images exhibit

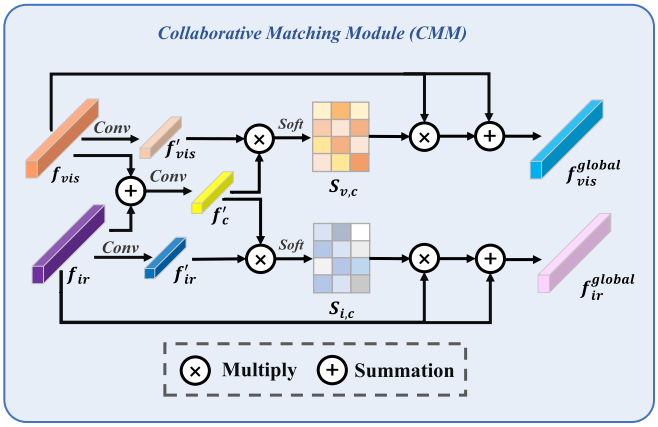

Fig. 4. Schematic diagram of collaborative matching module.

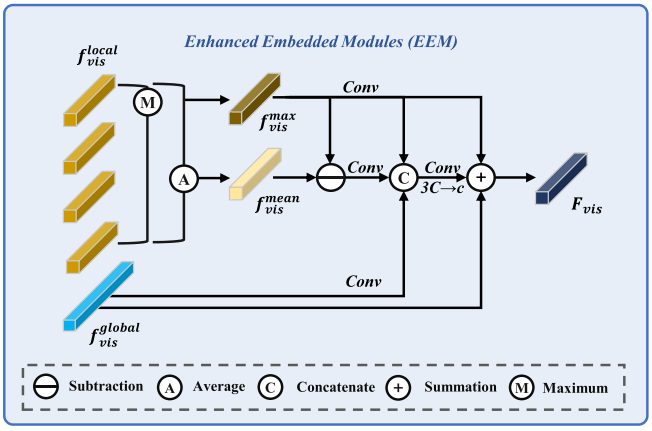

Fig. 5. Schematic diagram of enhanced embedded module.

robustness against variations in lighting and color, and they share similarities with infrared images in terms of shape, texture, and structural features. Leveraging these observations, we propose the integration of a Grey Foundation Module (GFM) that utilizes grayscale images to enhance the performance of visible images. The GFM module aims to enhance the discriminative nature of pedestrian contours and structural features while reducing the inclusion of redundant information that is not shared across modalities.

It is worth noting that the grey foundation module controls our training strategy. The whole training process is as follows: first, we generate the grey scale image x_g from the visible light image x_{vis}. next, the grey scale image is fed into the weight sharing Vit along with the infrared image x_{vis} to extract the features. Finally, the data is again replaced with visible and infrared images, which are fed together into the model. The above process can be expressed as:

$$f_{vis} = \mathcal{F}(f_{vis}),\ f_g = \mathcal{F}(f_g),\ f_{ir} = \mathcal{F}(f_{ir}), \tag{1}$$

f_{vis}, f_g and f_{ir} correspond to the extracted features of the visible, grey scale and infrared images respectively.

3.3. Collaborative matching module

After extracting features through the two-stream structure, we introduce the Collaborative Matching Module (CMM) to make full use of the complementary information between modalities. The CMM aims to guide the updating of modal features by adaptively aggregating the differences of the features. Specifically, we integrate all modal features using convolution and addition operations. Then, the softmax function is used to filter the noise points and obtain the matching information. Finally, to enhance the completeness of the features, we embed this information into the original features.

As depicted in Fig. 4, the visible and infrared features extracted from the backbone are fused to create an initial synergetic feature f_c. Meanwhile, in order to better get the relationship between the internal feature points, we introduce three 1×1 convolutions $\mathcal{G}_v$, $\mathcal{G}_i$, and $\mathcal{G}_c$ to deal with them, and obtain f'_{vis}, f'_{ir} and f'_c. Next, f'_c is individually matched with f'_{vis} and f'_{ir} using element-wise multiplication. To normalize the matching results, we apply the softmax function, resulting in $S_{i,c}$ and $S_{v,c}$, respectively. Finally, we enhance the initial features by embedding the matched information using multiplication and addition operations. The resulting global features for the visible and infrared branches are denoted as f_{vis}^{global} and f_{ir}^{global}, respectively. The entire process can be summarized as follows:

$$f'_c = \mathcal{G}_c(f_{vis} + f_{ir}), \tag{2}$$

$$S_{v,c} = Softmax[\mathcal{G}_v\ (f\ _{vis}) \times\ f'_c], \tag{3}$$

$$S_{i,c} = Softmax[\mathcal{G}_i\ (f\ _{ir}) \times\ f'_c], \tag{4}$$

$$f_{vis}^{global} = f_{vis} + f_{vis} \times S_{v,c}, \tag{5}$$

$$f_{ir}^{global} = f_{ir} + f_{ir} \times S_{i,c}, \tag{6}$$

Collaborative matching module enables the model to obtain compleme ntary information about each feature point within different modalities. Based on this information, the model is enabled to adaptively capture unique features and cues within the modality. It ultimately compensates for the limitations of each modality and achieves higher recognition accuracy.

3.4. Enhanced embedded module

Modality-specific variations (occlusion, viewing angle and illumination, etc.) also affect recognition accuracy. As shown in Fig. 5, we designed the enhanced embedding module aiming to improve the representation of intra-modal features and enhance the model's ability to suppress disturbances. It is worth noting that in the output of the Transformer decoder, we adopted the approach of TransReID [32] and enabled it to output four local features. Firstly, within the same modality, we compute the average and maximum values of all local feature points, denoted as f_{vis}^{max} and f_{vis}^{max} respectively. f_{vis}^{max} captures the most discriminative features, while f_{vis}^{mean} combines information from all features. However, it is worth noting that f_{vis}^{mean} is still susceptible to interference from background noise. Then, we subtract f_{vis}^{mean} from f_{vis}^{max} to get a transition feature f_{vis}^{trans}. f_{vis}^{trans} is used to optimise the feature description, which is a representation that filters the general information using discriminative information, and so can be robust to noise. Afterwards, we individually pass f_{vis}^{max}, f_{vis}^{trans}, and f_{vis}^{global} through a 1×1 convolutional layer and concatenate the resulting features to form f_{vis}^{con}. Finally, we integrate all the information to form the final feature F_{vis}. As follows:

$$F_{vis} = f_{vis}^{global} + \mathcal{G}_m(f_{vis}^{max}) + \mathcal{G}_{con}(f_{vis}^{con}), \tag{7}$$

Both $\mathcal{G}_m$ and $\mathcal{G}_{con}$ denote convolutional layers.

By leveraging the high discrimination nature of focal information, the model effectively corrects the comprehensive presentation of general information, thereby enhancing its robustness. Furthermore, the model's richness and completeness are improved by integrating the global wholeness and local uniqueness of the features. We apply the same operation to the infrared branch of the model, resulting in the extraction of the final visible feature, denoted as F_{vis}, and the infrared feature, denoted as F_{ir}. These features are concatenated to form the representation of the final pedestrian feature.

Table 1
Comparison with similar methods on two VI-ReID datasets (in %).

Methods	Venue	SYSU-MM01				RegDB			
		All-Search		Indoor-Search		Infrared to Visible		Visible to Infrared	
		Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
Two-stream [35]	ICCV2019	11.65	12.85	15.60	21.49	–	–	–	–
One-stream [35]	ICCV2019	12.04	13.67	16.94	22.95	–	–	–	–
Zero-Padding [35]	ICCV2019	14.80	15.95	20.58	26.92	16.70	17.90	17.80	18.90
JSIA-ReID [36]	AAAI2020	38.10	36.90	43.80	52.90	48.10	48.90	48.50	49.30
AlignGAN [35]	ICCV2019	42.40	40.70	45.90	54.30	56.30	53.40	57.90	53.60
AGW [1]	TPAMI2021	47.50	47.65	54.17	62.97	70.49	65.90	70.05	66.37
XIV-ReID [37]	AAAI2020	49.92	50.73	–	–	62.30	60.20	–	–
NFS [2]	CVPR2021	56.91	55.45	62.79	69.79	77.95	69.79	80.54	72.10
CoAL [38]	ACM MM2020	57.22	57.20	63.86	70.84	74.10	69.90	–	–
DG-VAE [39]	ACM MM2020	59.49	58.46	–	–	–	–	73.00	71.80
CM-NAS [40]	ICCV2021	61.99	60.02	67.01	72.95	82.57	78.31	84.54	80.32
SPOT [41]	TIP2022	65.34	62.25	69.42	74.63	79.37	72.26	80.35	72.46
SMCL [42]	ICCV2021	67.39	61.78	68.84	75.56	83.05	78.57	83.93	79.83
MPANet [43]	CVPR2022	70.58	68.24	76.74	80.95	82.80	80.70	83.70	80.90
FMCNet [44]	CVPR2022	66.34	62.51	68.15	74.09	88.38	83.86	89.12	84.43
PMT [34]	AAAI2023	67.53	64.98	71.66	76.52	84.16	75.13	84.83	76.55
DART [45]	CVPR2022	68.72	66.29	72.52	78.17	81.97	73.38	83.60	75.67
CM-EMD [15]	AAAI2023	73.39	68.56	80.53	82.71	92.77	86.85	94.37	88.23
Ours		**74.44**	**70.56**	**81.32**	**83.22**	**89.32**	**84.32**	**88.32**	**84.64**

3.5. Loss functions

In order to maximise the exploitation of different modal samples, we propose an adaptive triad loss. Specifically, we first randomly select m identities and n visible and infrared images in each identity. When grey scale and infrared images are input, the loss can be expressed as:

$$\mathcal{L}_{\text{gmax}} = \alpha \sum_{v=1}^{mn} [\max_{\forall y_v = y_k} \mathrm{D}\left(\mathrm{f}_{\mathrm{v}}^{\mathrm{g}}, \mathrm{f}_{\mathrm{k}}^{\mathrm{g}}\right) + 0.1], \tag{8}$$

$$\mathcal{L}_{\text{gmin}} = (1-\alpha) \sum_{v=1}^{mn} [\min_{\forall y_v \neq y_u} \mathrm{D}\left(\mathrm{f}_{\mathrm{v}}^{\mathrm{g}}, \mathrm{f}_{\mathrm{u}}^{\mathrm{g}}\right)], \tag{9}$$

$$\mathcal{L}_{\text{irmax}} = \alpha \sum_{v=1}^{mn} [\max_{\forall y_v = y_k} \mathrm{D}\left(\mathrm{f}_{\mathrm{v}}^{\mathrm{ir}}, \mathrm{f}_{\mathrm{k}}^{\mathrm{ir}}\right) + 0.1], \tag{10}$$

$$\mathcal{L}_{\text{irmin}} = (1-\alpha) \sum_{v=1}^{mn} [\min_{\forall y_v \neq y_u} \mathrm{D}\left(\mathrm{f}_{\mathrm{v}}^{\mathrm{ir}}, \mathrm{f}_{\mathrm{u}}^{\mathrm{ir}}\right)], \tag{11}$$

$$\mathcal{L}_{LOSS} = \mathcal{L}_{gmax} - \mathcal{L}_{gmin} + \mathcal{L}_{irmax} - \mathcal{L}_{irmin}, \tag{12}$$

where α is the adaptation factor, D (. .) denotes the distance metric, and y_v, y_u and y_k are the labels corresponding to the vth, uth and kth identities. Next, the loss is between-modalities when the visible and infrared images are input:

$$\mathcal{L}_{\text{ALT}} = \sum_{i=1}^{2PK} \left[\alpha \max_{\forall y_i = y_j} D\left(f_i, f_j\right) - (1-\alpha) \min_{\forall y_i \neq y_k} D\left(f_i, f_k\right) + 0.1 \right], \tag{13}$$

In addition to this loss, we also add the identity loss $\mathcal{L}_{\text{ID}}$ [33], modality-shared enhancement loss $\mathcal{L}_{\text{MSEL}}$ and discriminative center loss $\mathcal{L}_{\text{DCL}}$ in PMT [34] to explore the exploitation of more reliable information. The loss function can be defined as:

$$\mathcal{L}_{\text{overall}} = \mathcal{L}_{\text{ALT}} + 0.5\mathcal{L}_{\text{DCL}} + 0.5\mathcal{L}_{\text{MSEL}} + \mathcal{L}_{ID}. \tag{14}$$

4. Experiments

4.1. Datasets and evaluation metrics

SYSU-MM01 [12] is a large-scale VI-ReID dataset captured by four visible cameras and two infrared cameras. It comprises 286,628 visible images and 15,792 infrared images of 491 unique person identities. The training set contains 22,258 visible images and 11,909 infrared images, representing 395 person identities. The test set consists of 96 persons and includes 3803 infrared images for the query set, 301 single images randomly selected from the visible images, and 3010 multiple images for the gallery. Additionally, the dataset provides two search settings: indoor search, which employs indoor images, and full search, which utilizes all available images.

RegDB [13] is captured using 1 visible camera and 1 infrared camera. It comprises 8240 images of 412 unique person identities, with 206 identities used for training and 206 identities for testing. Each person in the dataset is represented by 10 infrared images and 10 visible images. Additionally, the dataset includes two evaluation settings: one for searching infrared images based on visible light images and another for searching visible light images based on infrared images. The evaluation process involves averaging the results over 10 iterations.

We extensively employ comprehensive evaluation metrics, namely the Cumulative Matching Characteristic (CMC) and the mean Average Precision (mAP). The CMC curve serves as an evaluation metric that measures the precision and recall of the model, while the mAP is an evaluation metric that assesses the overall quality of the model's comprehensive performance.

4.2. Implementation details

All experiments were conducted on an NVIDIA RTX3090 GPU. We utilized the Vit-B/16 model [23], pre-trained on ImageNet [46], as the backbone with a step size of 12.

To ensure consistency, the resolution of all images was adjusted to 256 × 128 pixels. During the training process, various data enhancement techniques were employed, such as random horizontal flipping, filling, random cropping, and random erasure operations. The training procedure involved the utilization of greyscale images as transition modalities during the initial phase of model training, in addition to the inclusion of infrared images. Subsequently, in the second training phase, both visible and infrared images were jointly fed into the model. No data augmentation was performed on the input images during testing. Our model employed a batch size of 40 and utilized up to 200 calendar elements, with the initial 20 calendar elements used for warm-up. Within each small batch, we randomly selected eight person identities and sampled four visible images and four infrared images for each identity. During the warm-up phase, the initial learning rate was set to 8×10^{-6} and gradually increased to 8×10^{-4}, which served as the starting learning rate for the remaining training process. The minimum learning rate was set to 8×10^{-6}. We fine-tuned the entire network using stochastic gradient descent (SGD) with a weight decay ratio of 1×10^{-4}.

Table 2
Performance comparison between different components (in %).

Methods				SYSU-MM01	
GFM	CMM	EEM	ATL	Rank-1	mAP
✓	×	×	×	64.68	62.44
✓	✓	×	×	67.26	65.36
✓	✓	✓	×	71.58	66.58
✓	✓	✓	✓	74.44	70.56

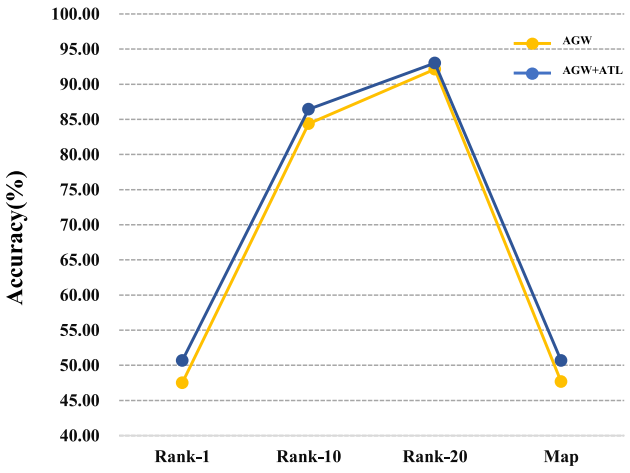

Fig. 6. Comparison results with CNN-based backbones.

After passing through the transformer layer, the feature dimension was reduced to 768. Following the example set by TransReID [32], we stochastically combined the local matrix to generate four local features. The optimal adaptation factor is 0.7.

4.3. Comparison with advanced methods

In this subsection, we compare the performance of TRR with other state-of-the-art methods. The methods compared include:

Results on SYSU-MM01 dataset. We evaluated the performance of TRR on the SYSU-MM01 [12] dataset, and the results are presented in Table 1. Under the full search setting, TRR achieved a Rank-1 accuracy of 74.44% and an mAP of 70.56%. This demonstrates a 1.05% improvement in Rank-1 accuracy and a 2.00% improvement in mAP compared to CM-EMD [15]. In the indoor search scenario, TRR demonstrated a Rank-1 accuracy improvement of 0.79% and an mAP improvement of 0.51% compared to CM-EMD [15].

Results on the RegDB dataset. We also conducted experiments on the RegDB dataset as shown in Table 1. We also performed experiments on the RegDB dataset as shown in Table 3. The Rank-1 accuracy of RCT in infrared to visible mode is 89.32% and mAP is 84.32%, and the Rank-1 accuracy in visible to infrared mode is 88.32% and mAP is 84.64%. RCT performs slightly lower than CM-EMD [15] on the RegDB dataset, but outperforms it on the SYSU-MM01 dataset. In fact, neither ResNet-based nor Vit-based methods can achieve the best performance on all datasets at the same time. Our TRR performs adaptive updating of each modal feature based on complementary information and distinctive cues between modalities. In order to enhance the expression of intra-modal features and the ability to suppress noise, the TRR extends and corrects the focal and integrative features within modalities, respectively. As a result, the proposed TRR is able to maintain high accuracy on multiple datasets at the same time and outperforms other methods in most cases.

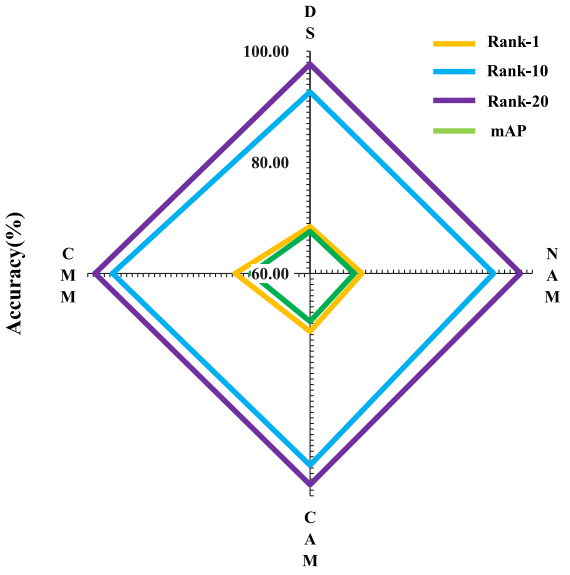

Fig. 7. Comparison of different strategies on the SYSU-MM01 dataset (in %).

4.4. Ablation study

In this section, we analyze the effects of the key modules, fusion strategies, and key parameters on the proposed method through an ablation study.

Effects of key components. To investigate the impact of each module on the model, a series of experiments was conducted on the SYSU-MM01 dataset. The experimental results are presented in Table 2, where × denotes deletion and ✓ denotes addition. A comparison between the first and second rows reveals that the addition of the collaborative matching module compensates for the limitations of the features by leveraging the complementary information across modalities. Consequently, the model achieves higher recognition accuracy. Similarly, incorporating enhanced embedding modules leads to an improvement in the model's performance. This observation highlights the feasibility of leveraging focal information to correct general information, and demonstrates how the fusion of global and local features enhances feature robustness. Lastly, optimizing sample selection through adaptive triple loss improves training results, further enhancing the model's performance. By conducting these experiments and analyzing the results, we provide evidence of the efficacy and benefits of each module in our proposed approach. These findings contribute to the understanding of pedestrian re-recognition in deep learning and validate the effectiveness of our method. Overall, the experiments conducted on the SYSU-MM01 dataset demonstrate the positive impact of the individual modules on the model's performance, establishing the significance and potential of our proposed approach.

It is worth noting that in order to further demonstrate the generality of the adaptive triplet loss. We conducted a comparison experiment on the cnn-based framework, choosing AGW [1] as an example. The experimental results are shown in Fig. 6, where we find that the addition of ATL improves the Rank-1 of the model by 3.97% and the mAP by 3.16%. This can prove that the adaptive ternary group loss, although effective on Vit, also has effectiveness in the CNN-based framework.

Impact of different strategies. The main challenge of this model is how to interact stably and efficiently with multiple features of different modalities in feature extraction. The collaborative matching module learns complementary information across modalities by adaptively associating and comparing feature points. To further demonstrate the effectiveness of the module, we introduced other strategies.

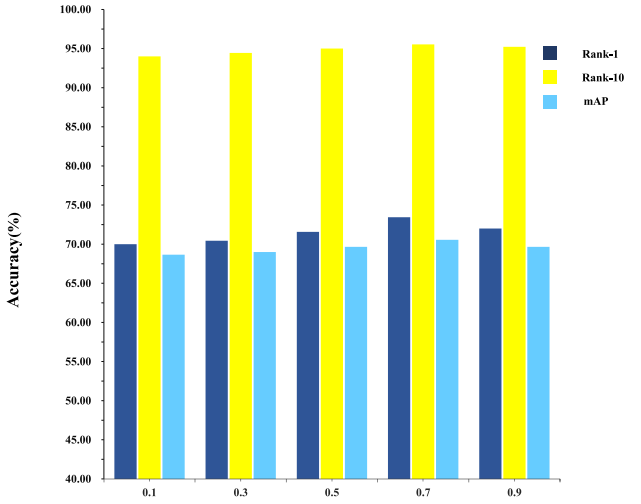

Fig. 8. Histogram of the effect of changes in α on VI-ReID.

The experimental results are presented in Fig. 7. The following operations are performed between features: the DS stands for the direct sum of features, the NLM stands for the Non-Local Attention Module [1], the CAM refers to the Cross Attention Module [47], and the CMM stands for the Collaborative Embedding Module.

Comparison of the results reveals that replacing these strategies does not significantly improve the model's performance. The direct summation of features yields the lowest accuracy. Moreover, the direct summation of features between modes is overly simplistic and crude, lacking the necessary flexibility and adaptability. Although the cross-attention mechanism performs relatively well, there is still room for improvement. Our analysis suggests that the cross-attention mechanism may overemphasize the interference information from a specific modality, leading to information redundancy and weakening the robustness of the features.

Key parameters. Next, we analyze the impact of the adaptation factor α in the adaptive triplet loss on the model's performance. The α in the formula controls the contribution of strong samples within the tuple. The optimal α can effectively find a balance between discriminating and similar samples. Considering samples with significant feature variations, while they facilitate model learning, the feature contents of infrared and visible images diverge more, potentially introducing irrelevant interference to the model and ultimately affecting its ability to perceive crucial information. We introduce α to alleviate the above problem. Fig. 8 illustrates the variation in model performance as α changes. As α ranges from 0.1 to 0.7, the model performance improves. The possible reason for this is that good comparison samples are adaptively selected from the sample set, allowing the model to achieve better training results. However, as α continues to increase, the comparison samples contain more and more noise information, and inaccurate features are amplified, affecting the robustness of the model.

5. Discussion

In this study, we present a multi-modal person re-identification (re-ID) method that utilizes transformer relation regularization. To begin, we employ a weight-sharing two-stream architecture to extract features from different modalities, while simultaneously enhancing the representation of structural and shape features through the incorporation of grey-scale images. Additionally, we introduce a collaborative matching module to effectively exploit the connected and complementary information between each modality. By adaptively capturing unique features and cues within each modality, we compensate for the limitations of individual modalities. Furthermore, the proposed enhanced embedding module not only improves the model's ability to suppress noise but also integrates discriminative and holistic information within each modality. To optimize the re-ID performance, we propose an adaptive triplet loss that offers advantages such as adaptive sample difficulty, effective utilization of sample relations, and improved feature generalization. The performance of our method is validated through extensive experiments conducted on a challenging VI-ReID dataset. Moving forward, we plan to investigate the incorporation of relatively novel techniques such as data augmentation and clustering to fully exploit the potential of our model.

CRediT authorship contribution statement

Xiangtian Zheng: Conceptualization, Writing – original draft, Editing, Investigation. **Xiaohua Huang:** Writing – review & editing, Supervision, Project administration, Methodology. **Chen Ji:** Formal analysis, Data curation, Writing – review & editing, Validation, Resources. **Xiaolin Yang:** Data curation, Writing — original draft. **Pengcheng Sha:** Conceptualization, Visualization, Project administration. **Liang Cheng:** Supervision, Project administration, Methodology, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgments

This work was supported by Nanjing Institute of Technology Scientific Research Start Fund (YKJ202118) and the Jiangsu Province Industry-University-Research Cooperation Project (DH20231709).

References

[1] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, S.C. Hoi, Deep learning for person re-identification: A survey and outlook, IEEE Trans. Pattern Anal. Mach. Intell. 44 (6) (2021) 2872–2893.
[2] Y. Chen, L. Wan, Z. Li, Q. Jing, Z. Sun, Neural feature search for rgb-infrared person re-identification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 587–597.
[3] Y. Gao, T. Liang, Y. Jin, X. Gu, W. Liu, Y. Li, C. Lang, MSO: Multi-feature space joint optimization network for rgb-infrared person re-identification, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 5257–5265.
[4] Z. Huang, J. Liu, L. Li, K. Zheng, Z.-J. Zha, Modality-adaptive mixup and invariant decomposition for RGB-infrared person re-identification, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, No. 1, 2022, pp. 1034–1042.
[5] Y. Zhu, Z. Yang, L. Wang, S. Zhao, X. Hu, D. Tao, Hetero-center loss for cross-modality person re-identification, Neurocomputing 386 (2020) 97–109.
[6] L. Zhang, G. Du, F. Liu, H. Tu, X. Shu, Global-local multiple granularity learning for cross-modality visible-infrared person reidentification, IEEE Trans. Neural Netw. Learn. Syst. (2021).
[7] H. Park, S. Lee, J. Lee, B. Ham, Learning by aligning: Visible-infrared person re-identification using cross-modal correspondences, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 12046–12055.
[8] X. Ning, K. Gong, W. Li, L. Zhang, X. Bai, S. Tian, Feature refinement and filter network for person re-identification, IEEE Trans. Circuits Syst. Video Technol. 31 (9) (2020) 3391–3402.
[9] X. Ning, K. Gong, W. Li, L. Zhang, JWSAA: joint weak saliency and attention aware for person re-identification, Neurocomputing 453 (2021) 801–811.
[10] E. Ning, C. Zhang, C. Wang, X. Ning, H. Chen, X. Bai, Pedestrian re-ID based on feature consistency and contrast enhancement, Displays (2023) 102467.
[11] M. Kim, S. Kim, J. Park, S. Park, K. Sohn, PartMix: Regularization strategy to learn part discovery for visible-infrared person re-identification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 18621–18632.

[12] A. Wu, W.-S. Zheng, H.-X. Yu, S. Gong, J. Lai, RGB-infrared cross-modality person re-identification, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5380–5389.
[13] D.T. Nguyen, H.G. Hong, K.W. Kim, K.R. Park, Person recognition system based on a combination of body images from visible light and thermal cameras, Sensors 17 (3) (2017) 605.
[14] H. Liu, D. Xia, W. Jiang, Towards homogeneous modality learning and multi-granularity information exploration for visible-infrared person re-identification, IEEE J. Sel. Top. Sign. Proces. (2023).
[15] Y. Ling, Z. Zhong, Z. Luo, F. Yang, D. Cao, Y. Lin, S. Li, N. Sebe, Cross-modality earth mover's distance for visible thermal person re-identification, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 37, No. 2, 2023, pp. 1631–1639.
[16] Q. Wu, J. Xia, P. Dai, Y. Zhou, Y. Wu, R. Ji, CycleTrans: Learning neutral yet discriminative features for visible-infrared person re-identification, 2022, arXiv preprint arXiv:2208.09844.
[17] Z. Zhao, B. Liu, Q. Chu, Y. Lu, N. Yu, Joint color-irrelevant consistency learning and identity-aware modality adaptation for visible-infrared cross modality person re-identification, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, No. 4, 2021, pp. 3520–3528.
[18] S. Shan, E. Xiong, X. Yuan, S. Wu, A knowledge-driven enhanced module for visible-infrared person re-identification, in: International Conference on Artificial Neural Networks, Springer, 2022, pp. 441–453.
[19] D. Cheng, X. Wang, N. Wang, Z. Wang, X. Wang, X. Gao, Cross-modality person re-identification with memory-based contrastive embedding, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 37, No. 1, 2023, pp. 425–432.
[20] Y. Zhang, Y. Yan, J. Li, H. Wang, MRCN: A novel modality restitution and compensation network for visible-infrared person re-identification, 2023, arXiv preprint arXiv:2303.14626.
[21] L. Wan, Z. Sun, Q. Jing, Y. Chen, L. Lu, Z. Li, G2DA: Geometry-guided dual-alignment learning for RGB-infrared person re-identification, Pattern Recognit. 135 (2023) 109150.
[22] Z. Wu, M. Ye, Unsupervised visible-infrared person re-identification via progressive graph matching and alternate learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 9548–9558.
[23] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, 2020, arXiv preprint arXiv:2010.11929.
[24] T. Wang, H. Liu, P. Song, T. Guo, W. Shi, Pose-guided feature disentangling for occluded person re-identification based on transformer, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, No. 3, 2022, pp. 2540–2549.
[25] L. Tan, P. Dai, R. Ji, Y. Wu, Dynamic prototype mask for occluded person re-identification, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 531–540.
[26] Y. Zhao, S. Zhu, D. Wang, Z. Liang, Short range correlation transformer for occluded person re-identification, Neural Comput. Appl. 34 (20) (2022) 17633–17645.
[27] H. Huang, A. Zheng, C. Li, R. He, et al., Parallel augmentation and dual enhancement for occluded person re-identification, 2022, arXiv preprint arXiv:2210.05438.
[28] Z. Ma, Y. Zhao, J. Li, Pose-guided inter-and intra-part relational transformer for occluded person re-identification, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 1487–1496.
[29] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[30] X. Ning, W. Tian, Z. Yu, W. Li, X. Bai, Y. Wang, HCFNN: high-order coverage function neural network for image classification, Pattern Recognit. 131 (2022) 108873.
[31] M. Jia, X. Cheng, S. Lu, J. Zhang, Learning disentangled representation implicitly via transformer for occluded person re-identification, IEEE Trans. Multimed. 25 (2022) 1294–1305.
[32] S. He, H. Luo, P. Wang, F. Wang, H. Li, W. Jiang, Transreid: Transformer-based object re-identification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 15013–15022.
[33] Z. Zheng, L. Zheng, Y. Yang, A discriminatively learned cnn embedding for person reidentification, acm Trans. Multimedia Comput. Commun. Appl. (TOMM) 14 (1) (2017) 1–20.
[34] H. Lu, X. Zou, P. Zhang, Learning progressive modality-shared transformers for effective visible-infrared person re-identification, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 37, No. 2, 2023, pp. 1835–1843.
[35] G. Wang, T. Zhang, J. Cheng, S. Liu, Y. Yang, Z. Hou, RGB-infrared cross-modality person re-identification via joint pixel and feature alignment, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 3623–3632.
[36] G.-A. Wang, T. Zhang, Y. Yang, J. Cheng, J. Chang, X. Liang, Z.-G. Hou, Cross-modality paired-images generation for RGB-infrared person re-identification, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, No. 07, 2020, pp. 12144–12151.
[37] D. Li, X. Wei, X. Hong, Y. Gong, Infrared-visible cross-modal person re-identification with an x modality, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, No. 04, 2020, pp. 4610–4617.
[38] X. Wei, D. Li, X. Hong, W. Ke, Y. Gong, Co-attentive lifting for infrared-visible person re-identification, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 1028–1037.
[39] N. Pu, W. Chen, Y. Liu, E.M. Bakker, M.S. Lew, Dual gaussian-based variational subspace disentanglement for visible-infrared person re-identification, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 2149–2158.
[40] C. Fu, Y. Hu, X. Wu, H. Shi, T. Mei, R. He, CM-NAS: Cross-modality neural architecture search for visible-infrared person re-identification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 11823–11832.
[41] C. Chen, M. Ye, M. Qi, J. Wu, J. Jiang, C.-W. Lin, Structure-aware positional transformer for visible-infrared person re-identification, IEEE Trans. Image Process. 31 (2022) 2352–2364.
[42] Z. Wei, X. Yang, N. Wang, X. Gao, Syncretic modality collaborative learning for visible infrared person re-identification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 225–234.
[43] Q. Wu, P. Dai, J. Chen, C.-W. Lin, Y. Wu, F. Huang, B. Zhong, R. Ji, Discover cross-modality nuances for visible-infrared person re-identification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 4330–4339.
[44] Q. Zhang, C. Lai, J. Liu, N. Huang, J. Han, FMCNET: Feature-level modality compensation for visible-infrared person re-identification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 7349–7358.
[45] M. Yang, Z. Huang, P. Hu, T. Li, J. Lv, X. Peng, Learning with twin noisy labels for visible-infrared person re-identification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 14308–14317.
[46] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, Ieee, 2009, pp. 248–255.
[47] J. Lu, D. Batra, D. Parikh, S. Lee, Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, Adv. Neural Inf. Process. Syst. 32 (2019).