Machine Vision and Applications (2023) 34:117
https://doi.org/10.1007/s00138-023-01473-4

ORIGINAL PAPER

A camera style-invariant learning and channel interaction enhancement fusion network for visible-infrared person re-identification

Haishun Du[1] · Xinxin Hao[1] · Yanfang Ye[1] · Linbing He[1] · Jiangtao Guo[1]

Received: 28 March 2023 / Revised: 15 September 2023 / Accepted: 18 September 2023 / Published online: 10 October 2023

Abstract
Cross-modality visible-infrared person re-identification (VI-ReID) aims to match visible and infrared pedestrian images from different cameras in various scenarios. However, most existing VI-ReID methods only focus on eliminating the modality discrepancy while ignoring the intra-class discrepancy caused by different camera styles. In addition, some feature fusion-based VI-ReID methods try to improve the discriminative capability of pedestrian representations by fusing pedestrian features from different convolutional layers or branches. However, most of them only implement feature fusion by simple operations, such as summation or concatenation, and ignore the interaction between different feature maps. To this end, we propose a camera style-invariant learning and channel interaction enhancement fusion network for VI-ReID. In particular, we design a channel interaction enhancement fusion module. It first computes and utilizes the channel-level similarity matrix of two feature maps to obtain two corresponding weighted feature maps that enhance the common concern information of the original two feature maps. Then, it obtains more discriminative pedestrian features by fusing the two weighted feature maps and mining their complementary information. Furthermore, in order to weaken the impact of camera style discrepancy of pedestrian images, we design a camera style-invariant feature-level adversarial learning strategy to ensure that the feature extraction network can extract camera style-invariant pedestrian features by the adversarial learning between the feature extraction network and the camera style classifier. Extensive experimental results on the two benchmark datasets, SYSU-MM01 and RegDB, demonstrate that the performance of CC-Net achieves the recent advanced level.

Keywords Cross-modality visible-infrared person re-identification · Channel interaction enhancement fusion · Camera style-invariant learning · Feature-level adversarial learning strategy

1 Introduction

Person re-identification (Re-ID), a technique that captures target persons from multiple cameras, has received a lot of attention due to its widespread use in security field [1–5]. In recent years, single-modality person re-identification [6–10] has achieved significant success. With the wide application of infrared cameras in night surveillance and low-light environment, how to match the images taken by visible cameras and infrared cameras has received a lot of attention from researchers [11–13]. However, in VI-ReID, pedestrian images not only have large intra-class discrepancy that is caused by different pedestrian poses [14–16] and different camera styles, but also have large modality discrepancy that is caused by the different reflection spectra of visible and infrared cameras [17]. So, VI-ReID is more challenging than single-modality person identification.

Convolutional neural networks have been widely used in the field of machine learning [18–21] due to the development of deep learning. At present, scholars have proposed many VI-ReID methods. Among them, the global and local feature-

✉ Haishun Du
jddhs@vip.henu.edu.cn

Xinxin Hao
haoxinxin@henu.edu.cn

Yanfang Ye
monica@henu.edu.cn

Linbing He
linbinghe@henu.edu.cn

Jiangtao Guo
jiangtaoguo@henu.edu.cn

[1] School of Artificial Intelligence, Henan University, Zhengzhou 450046, China

based VI-ReID methods are simple and effective [22–25]. These methods directly extract the global pedestrian features and the local pedestrian features, or fuse the global and local pedestrian features as the final pedestrian discriminative representations. However, most of them do not consider that the global features may contain a lot of background interference information, and the local features may have the problem of pedestrian misalignment. In addition, some scholars [15, 26, 27] tried to guide their networks to mine more discriminative pedestrian representations by fusing the pedestrian features from different convolutional layers or branches. However, most of these methods only achieve feature fusion by simple operations, such as summation or concatenation, and ignore the interaction between different feature maps, which hinders the improvement of person Re-ID model performance.

To ensure their networks to learn pedestrian representations that are robust to modality discrepancy, some scholars [28–30] used some cross-modality metric losses to supervise the training of their networks. In addition, some works [31–33] try to learn the modality-independent features by some adversarial learning strategies to weaken the impact of modality discrepancy. Specifically, most of these methods establish the connection between two modalities by using GANs to interconvert the pedestrian images of two modalities and then guide their networks to mine the pedestrian features that are more correlated with identities and less correlated with modalities. However, most existing methods only consider the modality discrepancy while ignoring the intra-class discrepancy between different images of a same pedestrian caused by camera style discrepancy, which results in insufficient discriminative capability of the pedestrian features extracted by these networks. It is worth pointing out that the camera style discrepancy mentioned in this paper refers to the differences in image styles caused by different camera viewpoints, different lighting conditions, and different backgrounds. As shown in Fig. 1, although both $Cam1$, $Cam2$, $Cam4$, and $Cam5$ are visible cameras, the styles of visible images captured by these cameras are markedly different due to the different shooting angles of $Cam1$ and $Cam4$, the different lighting conditions of $Cam1$ and $Cam5$, and the different shooting backgrounds of $Cam1$, $Cam4$, and $Cam5$. Similarly, although $Cam3$ and $Cam6$ are both infrared cameras, the styles of infrared images taken by them are also very different due to their different lighting conditions. We have found that the camera style discrepancy can significantly degrade the performance of VI-ReID models in our preliminary study. As a result, it is essential to design a strategy to ensure that VI-ReID models can extract camera style-invariant pedestrian features.

To solve the above-mentioned issues, we propose a novel network for VI-ReID and name it as camera style-invariant learning and channel interaction enhancement fusion network (CC-Net). Specifically, we first obtain two feature maps of a visible pedestrian image and two feature maps of an infrared pedestrian image using the backbone network, respectively. Then, we design a channel interaction enhancement fusion module (CIEFM), which first computes and utilizes the channel-level similarity matrix of two feature maps to obtain two corresponding weighted feature maps that enhance the common concern information of the original two feature maps and then obtains more discriminative pedestrian features by fusing the two weighted feature maps and mining their complementary information. Furthermore, to mitigate the impact of camera style discrepancy of pedestrian images, we design a camera style-invariant feature-level adversarial learning strategy that enables the feature extraction network to extract camera style-invariant pedestrian features by the adversarial learning between the feature extraction network and the camera style classifier. Extensive experimental results on the two benchmark datasets, SYSU-MM01 and RegDB, demonstrate that CC-Net can effectively improve the performance of VI-ReID.

Fig. 1 Pedestrian images in the SYSU-MM01 dataset. Each line contains different images of a same person taken by six cameras (Cam1, Cam2, Cam4, and Cam5 are visible cameras, while Cam3 and Cam6 are infrared cameras)

The major contributions of this paper are summarized as follows:

(1) We propose an end-to-end CC-Net to extract camera style-invariant discriminative pedestrian features for VI-ReID.
(2) We design a channel interaction enhancement fusion module (CIEFM) to obtain more discriminative pedes-

trian features by enhancing the common concern information of two feature maps and mining their complementary information.
(3) We design a camera style-invariant feature-level adversarial learning strategy to weaken the impact of camera style discrepancy of pedestrian images on the performance of VI-ReID.
(4) Extensive experimental results on the two benchmark datasets, SYSU-MM01 and RegDB, demonstrate that the performance of CC-Net achieves the recent advanced level.

2 Related works

2.1 Single-modality person Re-ID

For the past few years, some scholars have presented a number of effective methods [34–39] to solve cross-camera matching problems of single-modality pedestrian images. Based on the loss functions that they use, these methods can be classified as the representation-learning-based methods [36] and the metric-learning-based methods [39]. Among them, the representation-learning-based methods are frequently used for person Re-ID tasks. They usually use ID losses or attribute losses to train their networks. For example, Zhang et al. [35] proposed an IDE network and used an ID loss to supervise the training of their network. Lin et al. [36] considered that only using an ID loss to train networks is unable to capture more discriminative pedestrian representations. To solve this problem, they used an ID loss and an attribute loss to jointly train their network, which enables their network to learn more discriminative pedestrian representations. The metric-learning-based methods aim to optimize the relative distances between different images using metric losses so that two images with a same identity have a relatively smaller distance than two images with different identities. For example, Cheng et al. [37] used an improved triplet loss to train their network so that the positive sample pairs have a relatively smaller distance than the positive–negative sample pairs. Hermans et al. [38] proposed a hard sample sampling triplet loss (TriHard loss), which uses harder samples to train their network so that their network has a strong generalization capability. Although the above methods can solve visible pedestrian image matching problems well, they perform poorly in solving the cross-modality pedestrian image matching problems.

2.2 Cross-modality person Re-ID

Cross-modality person Re-ID aims to match pedestrian images captured by different kinds of cameras, such as visible cameras and infrared cameras [40–42]. However, since there are significant discrepancies between different modality images, VI-ReID requires addressing not only the challenges of pose variations, camera viewpoint variations, occlusions, and cluttered backgrounds, but also the modality discrepancy. Some researchers have tried to use feature-level-based and image-level-based methods to learn the modality-shared information of pedestrian images in different modalities [12, 43–45], or to design some new loss functions [28, 46, 47] to weaken the impact of modality discrepancy on the overall performance of networks. For example, Ye et al. [44] presented a MACE learning method, which can address intra- and inter-modality variations at both the feature and classifier layers. Liu et al. [46] presented a hetero-center triplet loss, which enables their network to capture more discriminative pedestrian representations by calculating the distances between the anchor class feature centers and the positive/negative class feature centers. Choi et al. [48] separated ID-discriminative factors and ID-excluded factors from cross-modality images and then combined them to generate modality-different but identity-consistent images. Zhang et al. [49] presented a dual-path cross-modality feature learning framework that takes the inherent spatial structure and the discrepancy between cross-modality image pairs into account. Wan et al. [50] proposed a geometrically guided dual alignment learning method that weakens the discrepancy between the two modalities by converting RGB images and IR images into semantically aligned images. Sun et al. [51] proposed a CAA strategy that reduces the discrepancy between two modalities by mining intra-modality attentional information with counter-factual causality. In addition, some adversarial learning strategies have been used to handle the problems in VI-ReID [31, 32, 52, 53]. Specifically, most of these methods reduce the modality discrepancy at the image level by transforming the pedestrian images of two modalities to each other. However, most existing VI-ReID methods only focus on eliminating the modality discrepancy between pedestrian images of two modalities, while ignoring the camera style discrepancy, which can significantly affect the performance of VI-ReID. To this end, we design a camera style-invariant feature-level adversarial learning strategy, which enables the feature extraction network to have a certain capability of extracting camera style-invariant pedestrian features by the adversarial learning between the feature extraction network and the camera style classifier.

2.3 Feature fusion-based person Re-ID

Feature fusion achieves promising achievements in image semantic segmentation [54], face recognition [55], etc. Recently, some scholars [26, 27] have introduced feature fusion into the field of person Re-ID to improve the discriminative capability of pedestrian representations. For example, Zhao et al. [15] guided their network to mine pedestrian

representations with robustness and discriminative capability by fusing regional features at different semantic levels using a tree structure feature fusion strategy. Xiang et al. [27] proposed a deep multi-modality fusion network (DMF), which significantly enhances the generalization capability of models by introducing rich semantic knowledge and multi-modality fusion strategy. Liu et al. [26] improved the discriminative capability and robustness of pedestrian features by fusing the output features of the middle layer with the final output features of the backbone network. However, most existing feature fusion-based methods achieve feature fusion only by simple operations, such as summation or concatenation, while ignoring the interaction between different feature maps, which limits the improvement of the performance of VI-ReID. To this end, we design a channel interaction enhancement fusion module (CIEFM) to obtain more discriminative pedestrian representations by enhancing the common concern information of two feature maps and mining their complementary information.

3 Proposed method

The basic framework of CC-Net is shown in Fig. 2. CC-Net is mainly made up of a two-stream backbone network (ResNet-50), two channel interaction enhancement fusion modules (CIEFMs) and a camera style classifier. Among them, the two-stream backbone network and the two channel interaction enhancement fusion modules together form the feature extraction network M. Specifically, the first branch of the two-stream backbone network is used to obtain the feature maps of visible images, and the second branch is used to obtain the feature maps of infrared images. The two channel interaction enhancement fusion modules are used to fuse the two feature maps output from each of the two branches of the backbone network, respectively, which can extract more discriminative pedestrian features. Specifically, we utilize a camera style-invariant feature-level adversarial learning strategy to perform the adversarial learning between the feature extraction network M and the camera style classifier W_C, which finally enables the feature extraction network M to extract camera style-invariant pedestrian features.

3.1 Two-stream backbone network

The two-stream backbone network contains a visible branch and an infrared branch, as shown in Fig. 2. In the two branches, the parameters of *S*tage1 of each branch are specific to capture the modality-specific information, while the parameters of *S*tage2 $\sim$ *S*tage5 are shared to capture the modality-shared information. In addition, we modify *S*tage5 of each branch into two convolutional blocks, which have the same structure but different parameters.

We feed visible pedestrian images and infrared pedestrian images into the corresponding branches of the backbone network. For a visible pedestrian image x^v, the two feature maps ${F_1}^v$ and ${F_2}^v$ output from the visible branch of the backbone network, which are represented as follows:

$$\begin{cases} {F_1}^v = \phi_1\left(f\left(x^v\right)\right) \\ {F_2}^v = \phi_2\left(f\left(x^v\right)\right) \end{cases}, \tag{1}$$

where $f(\cdot)$ denotes *S*tage1 $\sim$ *S*tage4 in the visible branch, $\phi_1(\cdot)$ and $\phi_2(\cdot)$ denote the two convolutional blocks of *S*tage5 in the visible branch. Similarly, for an infrared pedestrian image x^t, the two feature maps ${F_1}^t$ and ${F_2}^t$ can be obtained after processing by the infrared branch of the backbone network.

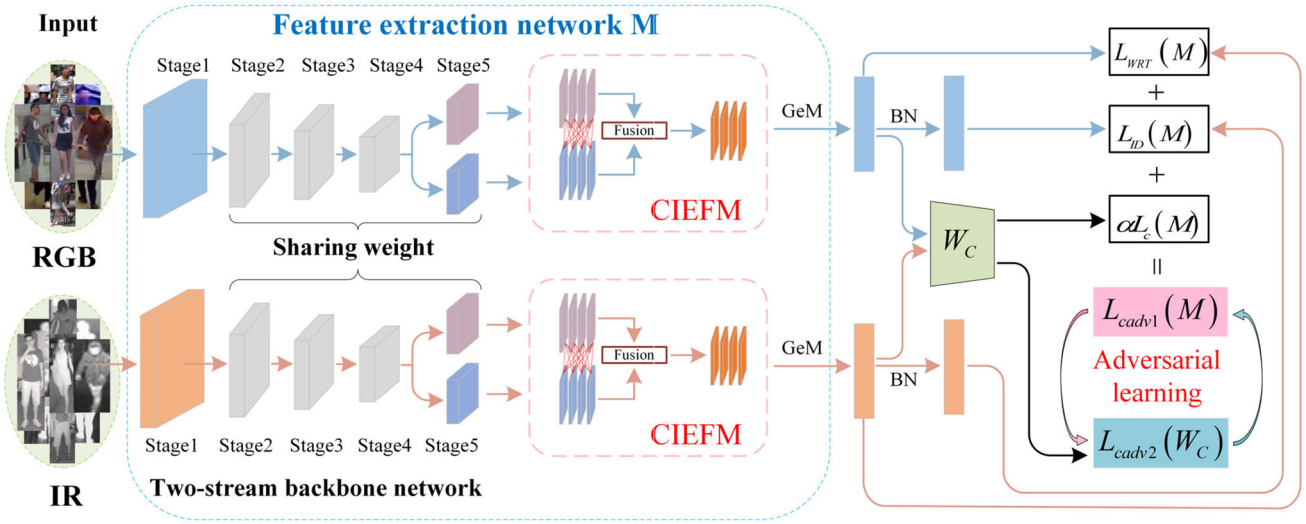

Fig. 2 The basic framework of CC-Net

3.2 Baseline model

The baseline model (baseline) is mainly made up of ResNet-50, ID loss, WRT loss, and GeM pooling. Specifically, we use a two-stream network with ResNet-50 as the backbone to process images of two modalities. In particular, for the first convolutional block, the two branches of the dual-stream network use the same structure but different parameters, with the aim of learning modality-specific features of visible and infrared images, respectively. For the remaining four convolutional blocks, the two branches of the dual-stream network share weights, aiming to extract modality-invariant features of visible and infrared images. In addition, we use ID loss L_{ID} and WRT loss L_{WRT} [2] as baseline loss L_{Base}, i.e.,

$$L_{Base} = L_{ID} + L_{WRT}. \tag{2}$$

3.3 Channel interaction enhancement fusion module

Most existing feature fusion-based person Re-ID methods usually use simple operations, such as summation or concatenation, to achieve feature fusion, and do not consider the interaction between different feature maps. Therefore, as shown in Fig. 3, we design a channel interaction enhancement fusion module (CIEFM), which obtains more discriminative pedestrian features by enhancing the common concern information of two feature maps and mining their complementary information. Specifically, the module first computes and utilizes the channel-level similarity matrix of two feature maps to obtain two corresponding weighted feature maps that enhance the common concern information of the original two feature maps. Then, it obtains more discriminative pedestrian features by fusing the two weighted feature maps and mining their complementary information.

Given two feature maps $F_1 \in \mathbb{R}^{C\times H\times W}$ and $F_2 \in \mathbb{R}^{C\times H\times W}$, in which C, H, and W denote the channel, the height, and the width of the feature maps, respectively. We first reshape the feature map F_1 and the feature map F_2 into $\widetilde{F}_1 \in \mathbb{R}^{C\times l}$ and $\widetilde{F}_2 \in \mathbb{R}^{C\times l}$, respectively, where $l = H \times W$. Then, we obtain the channel-level similarity matrix M of the two feature maps by performing bilinear operation on $\widetilde{F}_1$ and $\widetilde{F}_2$. Finally, based on the similarity matrix M, we calculate the weight matrix W using the following formula:

$$W_{ij} = \frac{\exp(-M_{ij})}{\sum_{k=1}^{C}\exp(-M_{ik})}, \tag{3}$$

where i denotes the ith channel of feature map F_1, and j denotes the jth channel of feature map F_2.

We use the weight matrix W to enhance the common concern information in F_1 and F_2, respectively, and then obtain the corresponding weighted feature maps $F_{W1} \in \mathbb{R}^{C\times H\times W}$ and $F_{W2} \in \mathbb{R}^{C\times H\times W}$ as follows:

$$\begin{cases} F_{W1} = reshape(W \times \widetilde{F}_1) \\ F_{W2} = reshape(W \times \widetilde{F}_2) \end{cases}. \tag{4}$$

In addition, considering that the complementary information between feature maps is essential to enhance the discriminative capability of pedestrian features, we further fuse the weighted feature maps F_{W1} and F_{W2} to obtain the final output feature map $F \in \mathbb{R}^{C\times H\times W}$ as follows:

$$F = F' \otimes F_{W1} + \left(1 - F'\right) \otimes F_{W2}, \tag{5}$$

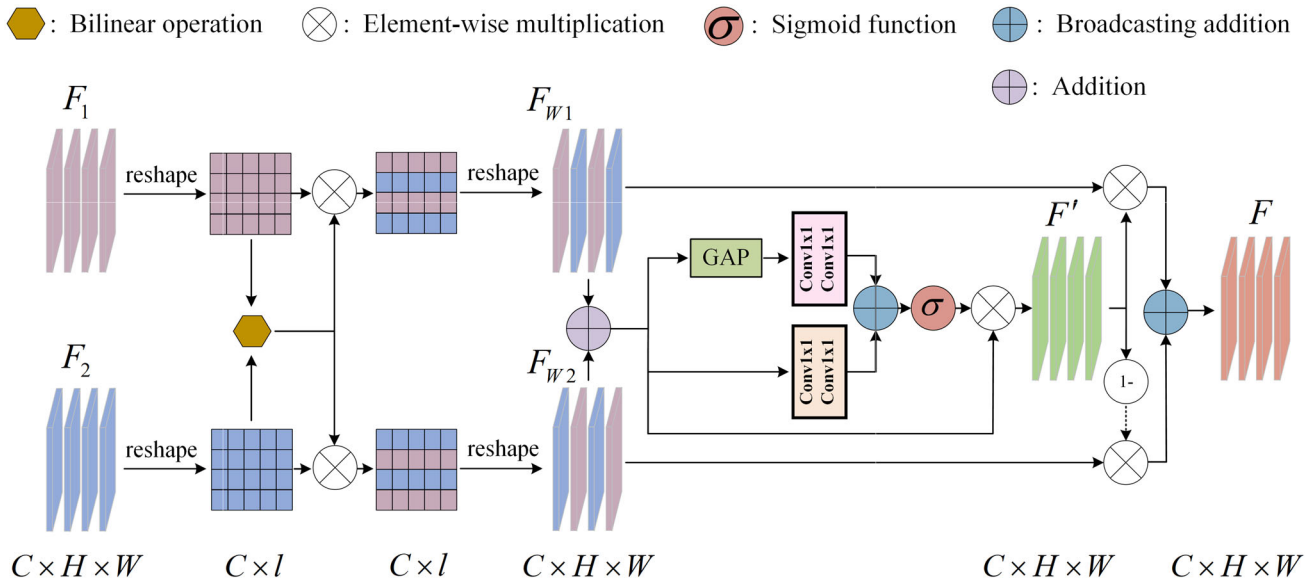

Fig. 3 The architecture of CIEFM

where F' denotes the fusion weight map, which is calculated as follows:

$$F' = (F_{W1} \oplus F_{W2}) \otimes \sigma\left(\varphi_1\left(GAP\left(F_{W1} \oplus F_{W2}\right)\right) + \varphi_2\left(F_{W1} \oplus F_{W2}\right)\right), \tag{6}$$

where $\varphi_1(\cdot)$ and $\varphi_2(\cdot)$ denote two convolutional blocks, GAP$(\cdot)$ denotes the global average pooling operation, and $\sigma\{\cdot\}$ denotes the Sigmoid function. It should be noted that instead of adding F_{W1} and F_{W2} directly to get F', we further process the result of adding F_{W1} and F_{W2} to better aggregate multi-scale contextual information along the channel dimension, which allows our network to concatenation on both global and local discriminative information contained in the two feature maps.

3.4 Camera style-invariant feature-level adversarial learning strategy

Algorithm 1 The detailed training pipeline of the camera style-invariant feature-level adversarial learning strategy

Input: Training images with camera information and ID labels
$X = \left\{x_i^c \,|i = 1, 2, \cdots m; c = 1, 2, \cdots n\right\}$.
Total number of training epochs max _$epochs$.
Training batch size $Batchsize$.
Threshold T.

Initialization: The feature extraction network M.
The camera style classifier W_C.
$epoch = 1$.

Output: The feature extraction network M.

1. **while** $epoch \leq$ max _$epoches$ **do**:
2. Randomly select Batchsize pedestrian images from X using the PK sample strategy and extract the corresponding pedestrian features using the feature extraction network M;
3. **if** epoch < T **do**:
4. Make $\alpha = 0$ in formula (8) and calculate the loss $L_{cadv1}(M)$ using formula (8);
5. Update the parameters of the feature extraction network M using the BP algorithm;
6. Calculate the loss $L_{cadv2}(W_C)$ using formula (9);
7. Update the parameters of the camera style classifier W_C using the BP algorithm.
8. **else do**:
9. Calculate the loss $L_{cadv2}(W_C)$ using formula (9);
10. Update the parameters of the camera style classifier W_C using the BP algorithm;
11. Make $\alpha = 1$ in formula (8) and calculate the loss $L_{cadv1}(M)$ using formula (8);
12. Update the parameters of the feature extraction network M using the BP algorithm.
13. **end if**
14. epoch = epoch +1.
15. **end while**

Pedestrian images taken by different cameras vary greatly in camera style due to differences in camera viewpoints, lighting conditions and backgrounds, and the camera style discrepancy can significantly affect the overall performance of networks. Some scholars [31–33] weakened the impact of camera style discrepancy on person Re-ID performance using GANs to transform pedestrian images with different camera styles to each other. However, using GANs to generate images not only requires huge computational resources, but also may introduce additional noise.

To further weaken the impact of camera style discrepancy of pedestrian images, we design a camera style-invariant feature-level adversarial learning strategy. Specifically, we introduce a camera style classifier W_C. By the adversarial learning between the camera style classifier W_C and the feature extraction network M, the feature extraction network M has a certain capability of extracting camera style-invariant pedestrian features. In particular, the outputs of the camera style classifier W_C are the probabilities of a pedestrian image belonging to the ith camera style C_i $(i = 1, 2, \cdots, n)$ and uniform style C_{n+1}. In order to enable the feature extraction network M to extract camera style-invariant pedestrian features, we should optimize the feature extraction network M so that W_C can classify all the pedestrian image features extracted by the feature extraction network M into a uniform style C_{n+1}, i.e., optimizing the following loss:

$$L_c(M) = \psi_{ce}\left(W_C\left(M(x)\right), C_{n+1}\right), \tag{7}$$

where x denotes the pedestrian image, and $\psi_{ce}(\cdot)$ denotes the cross-entropy loss. In addition, we also introduce the weighted regularization triplet (WRT) loss [2] and the ID loss to enhance the identity-related information in pedestrian features. In summary, we need to optimize the following loss:

$$L_{cadv1}(M) = L_{ID} + L_{WRT} + \alpha L_c, \tag{8}$$

where L_{WRT} and L_{ID} denote the WRT loss and the identity-related identity (ID) loss, respectively. In particular, we use the cross-entropy loss as the ID loss.

To enhance the classification capability of the camera style classifier W_C as much as possible, we also need to optimize the following loss:

$$L_{cadv2}(W_C) = \psi_{ce}\left(W_C\left(M(x)\right), C_i\right), i = 1, 2, 3, \cdots n. \tag{9}$$

In the training phase, we use two optimizers to independently optimize L_{cadv1} and L_{cadv2} to achieve the adversarial learning between the camera style classifier W_C and the feature extraction network M. Algorithm 1 shows the detailed training process of the camera style-invariant feature-level adversarial learning. As can be seen from Algorithm 1, we use

a threshold T to control when $L_c(M)$ participates in network training. When epoch $<$ T, we set $\alpha = 0$ in $L_{cadv1}(M)$. At this time, $L_{cadv1}(M)$ and $L_{cadv2}(W_C)$ supervise network training together, with the purpose of training a better performing feature extraction network M and a better performing camera style classifier W_C. When epoch $\geq$ T, we set $\alpha = 1$ in $L_{cadv1}(M)$ to introduce $L_c(M)$ in $L_{cadv1}(M)$. In this case, the feature extraction network M and the camera style classifier W_C begin adversarial learning. With the gradual updating of the feature extraction network M and the camera style classifier W_C , although W_C has a strong capability to classify camera styles, the pedestrian features extracted by the feature extraction network M are still classified into a uniform style C_{n+1}. At this point, we think that the feature extraction network M has the capability to exact camera style-invariant pedestrian features.

4 Experiments

4.1 Datasets and experimental settings

4.1.1 Datasets and evaluation metric

We evaluate the overall performance of CC-Net on the two benchmark datasets, RegDB [56] and SYSU-MM01 [11], respectively. The detailed information of the two benchmark datasets is listed in Table 1.

SYSU-MM01 dataset is the first publicly available large-scale dataset for VI-ReID provided by Sun Yat-sen University. The images in this dataset were captured by a total of 6 cameras, which contains four visible cameras and two infrared cameras. A total of 30,071 visible images and 15,792 infrared images of 491 pedestrians are included in this dataset. According to the evaluation protocol in [11], we randomly select 34,167 images of 395 pedestrians from this dataset to form the training set, and 4104 images of the remaining 96 pedestrians to form the testing set. In addition, we adopt two evaluation modes, i.e., all-search mode and indoor-search mode. For the all-search mode, all images captured by the six cameras are used. For the indoor-search mode, only the indoor images captured by the first, second, third, and sixth cameras are used.

RegDB dataset is a publicly available dataset for VI-ReID provided by Dongguk University in Korea. This dataset contains 8240 images of 412 pedestrians taken by 2 cameras, in which 254 are females and 158 are males. According to the evaluation protocol in [29], we randomly select 4120 images of 206 pedestrians from this dataset as the training set and 4120 images of the remaining 206 pedestrians as the testing set. In the testing phase, we adopt two test modes, i.e., visible-to-infrared mode and infrared-to-visible mode.

Evaluation metric. In this paper, all experiments use the cumulative matching characteristics (CMCs) and the mean average precision (mAP) as the metric of VI-ReID performance. The mAP considers both accuracy and completeness to assess the overall performance of the experimental results. It reflects the degree to which the correctly identified pedestrian images are ahead of the other retrieval results.

4.1.2 Implementation details

We use the ResNet-50 [57] pretrained on the ImageNet [58] as the backbone network. In particular, we change the original stride size 2 to 1 in the two residual blocks of the *S*tage5 in each branch of the two-stream backbone network to obtain more rich pedestrian features. In the training phase, we use data augmentation techniques such as horizontal flipping, random erasing, random cropping, and random channel exchange. In addition, the size of all pedestrian images is resized to 288 $\times$ 144, and the batchsize is set to 64. We use two SGD optimizers with momentum of 0.9 and weight decay of 5×10^{-4} to optimize the parameters of the feature extraction network M and the camera style classifier W_C with a total of 200 epochs, respectively. Moreover, we initialize the learning rate to 0.1, which decays to 0.01 and 0.001 at the 20th and 50th epochs, respectively. It is worth mentioning that, for the threshold T in Algorithm 1, we set it to 60 in our experiments.

4.2 Comparison with some state-of-the-art methods

The experimental results of CC-Net and some state-of-the-art methods on the SYSU-MM01 and RegDB datasets are shown in Tables 2 and 3, respectively.

Evaluation on SYSU-MM01. Table 2 shows that CC-Net achieves 67.74% Rank-1 accuracy and 62.81% mAP in the all-search mode, and 73.85% Rank-1 accuracy and 77.42% mAP in the indoor-search mode. In the two search modes, compared with the advanced VI-ReID method SMCL [59], the Rank-1 accuracy of CC-Net is improved by 0.35% and 5.01%, and the mAP is improved by 1.03% and 1.86%,

Table 1 The detailed information of the two benchmark datasets

Dataset	Cameras	Training set (IDs/images)	Testing (IDs/images)
SYSU-MM01	6	395/34,167	96/4104
RegDB	2	206/4120	206/4120

Table 2 Comparison with some state-of-the-art methods on the SYSU-MM01 dataset

Settings	All-search				Indoor search			
Method	Rank-1	Rank-10	Rank-20	mAP	Rank-1	Rank-10	Rank-20	mAP
Zero-Pad [11]	14.80	54.12	71.33	15.95	20.58	68.38	85.79	26.92
HCML [29]	14.32	53.16	69.17	16.16	24.52	73.25	86.73	30.08
cmGAN [31]	26.97	67.51	80.56	31.49	31.63	77.23	89.18	42.19
MAC [60]	33.26	79.04	90.09	36.22	36.43	62.36	71.63	37.03
AlignGAN [52]	42.40	85.00	93.70	40.70	45.90	87.60	94.40	54.30
BDTR [61]	27.32	66.96	81.07	27.32	32.46	77.42	89.62	42.46
DGD+MSR [30]	37.35	83.40	93.34	38.11	39.64	89.29	97.66	50.88
Xmodal [62]	49.92	89.79	95.96	50.73	–	–	–	–
DDAG [43]	54.75	90.39	95.81	53.02	61.02	94.06	98.41	67.98
JSIA-ReID [63]	38.10	80.70	89.90	36.90	43.80	86.20	94.20	52.90
LZM [64]	45.00	89.06	–	45.94	49.66	92.47	-	59.81
MSPAC [65]	46.62	87.59	95.77	47.26	51.63	93.48	98.82	61.54
DTCL [66]	54.14	89.93	96.13	54.14	–	–	–	–
HAT [67]	55.29	92.14	97.36	53.89	62.10	95.75	99.20	69.37
cm-SSFT [68]	61.60	89.20	93.90	**63.20**	70.50	94.90	97.70	72.60
HCT [46]	61.68	93.10	97.17	57.51	63.41	91.96	95.28	68.17
AGW [2]	47.50	84.39	92.14	47.65	54.17	91.14	95.98	62.97
SMCL [59]	67.39	92.87	96.76	61.78	68.84	96.55	98.77	75.56
G^2DA [50]	63.94	93.34	97.29	60.73	71.06	**97.31**	**99.47**	76.01
CAA [51]	59.46	91.84	96.75	58.83	65.23	96.64	99.21	71.42
CC-Net	**67.74**	**94.61**	**97.96**	62.81	**73.85**	96.57	99.11	**77.42**

"–" no available data
Bold values indicate the best performance

Table 3 Comparison with some state-of-the-art methods on the RegDB dataset

Settings	Visible-to-infrared				Infrared-to-visible			
Method	Rank-1	Rank-10	Rank-20	mAP	Rank-1	Rank-10	Rank-20	mAP
Zero-Pad [11]	17.75	34.21	44.35	18.90	16.63	34.68	44.25	17.82
HCML [29]	24.44	47.53	56.78	20.08	21.7	45.02	55.58	22.24
MAC [60]	36.43	62.36	71.63	37.03	36.2	61.68	70.99	36.63
BDTR [61]	33.56	58.61	67.43	32.76	32.92	58.46	68.43	31.96
D^2RL [33]	43.40	66.10	76.30	44.10	–	–	–	–
DDAG [43]	69.34	86.19	91.49	63.46	68.06	85.15	90.31	61.8
LZM [64]	57.03	76.10	84.34	58.06	–	–	–	–
Xmodal [62]	62.21	83.13	91.72	60.18	–	–	–	–
NFS [70]	80.54	91.96	95.07	72.1	–	–	–	–
HAT [67]	71.83	87.16	92.16	67.57	70.02	86.45	91.96	66.30
AGW [2]	70.05	86.21	91.55	66.37	70.49	87.12	91.84	65.90
cm-SSFT [68]	72.30	–	–	72.90	71.00	–	–	71.70
MCLNet [69]	80.31	92.70	96.03	73.07	75.93	90.93	94.59	69.49
G^2DA [50]	73.95	89.47	93.67	65.49	69.67	86.41	91.38	61.98
CAA [51]	80.31	92.45	96.12	73.54	79.87	92.23	**95.91**	72.36
CC-Net	**87.09**	**95.80**	**97.57**	**76.73**	**82.55**	**92.59**	95.31	**72.48**

"–" no available data
Bold values indicate the best performance

respectively. In the two search modes, compared with the advanced VI-ReID method G^2DA [50], the Rank-1 accuracy of CC-Net is improved by 3.8% and 2.79%, and the mAP is improved by 2.08% and 1.41%, respectively. In the two search modes, compared with the advanced VI-ReID method CAA [51], the Rank-1 accuracy of CC-Net is improved by 8.28% and 8.62%, and the mAP is improved by 3.98% and 6%, respectively. These results demonstrate that the performance of CC-Net achieves the recent advanced level.

Evaluation on RegDB. Table 3 shows that CC-Net achieves 87.09% Rank-1 accuracy and 76.73% mAP in the visible-to-infrared mode, and 82.55% Rank-1 accuracy and 72.48% mAP in the infrared-to-visible mode. In the two search modes, compared with the advanced VI-ReID method MCLNet [69], the Rank-1 accuracy of CC-Net is improved by 6.78% and 6.62%, and the mAP is improved by 3.66% and 2.99%, respectively. In the two search modes, compared with the advanced VI-ReID method G^2DA [50], the Rank-1 accuracy of CC-Net is improved by 13.14% and 12.88%, and the mAP is improved by 11.24% and 10.05%, respectively. In the two search modes, compared with the advanced VI-ReID method CAA [51], the Rank-1 accuracy of CC-Net is improved by 6.78% and 2.68%, and the mAP is improved by 3.19% and 0.12%, respectively. These results show again that our method has some advantages compared with most advanced VI-ReID methods.

However, as can be seen from Table 2, in the indoor-search mode, the Rank-10 and Rank-20 of CC-Net decrease 0.74% and 0.36%, respectively, compared with the Rank-10 and Rank-20 of G^2DA. As can be seen from Table 3, in the infrared-to-visible mode, the Rank-20 of CC-Net is 0.6% lower than that of CAA. This may be due to the fact that our model cannot completely eliminate the modality discrepancy caused by pixel-level changes in the images.

4.3 Computational complexity

In order to give an idea of the computational complexity of our model, we conduct an experimental study on the SYSU-MM01 dataset. The training time of our model is about 5.6 h. Although it takes a long time to train our CC-Net, we can train it offline and use it online in real-world applications, and its online inference time for a single image is only 0.008 s.

4.4 Ablation experiments

We conduct ablation experiments on the SYSU-MM01 and RegDB datasets, respectively, to evaluate the performance of different modules of CC-Net. Specifically, the baseline model (Baseline) is mainly made up of ResNet-50, ID loss, WRT loss, and GeM pooling. We construct Baseline + CIEFM, Baseline + CSIL, and Baseline + CIEFM + CSIL to demonstrate the performance of different modules of CC-Net. In particular, CSIL denotes the camera style-invariant feature-level adversarial learning strategy. The ablation experimental results are listed in Table 4. From Table 4, we can find that, on the two benchmark datasets, compared with the Baseline, Baseline + CIEFM achieves improvements of 2.93% and 4.12% in Rank-1 accuracy, and 2.24% and 2.58% in mAP, respectively. We can see from these results that CIEFM can extract more discriminative pedestrian representations, which effectively increases the performance of Baseline. On the two benchmark datasets, compared with the Baseline, Baseline + CSIL achieves improvements of 2.50% and 5.14% in Rank-1 accuracy, and 2.28% and 2.30% in mAP, respectively. We can see from these results that training Baseline using the camera style-invariant feature-level adversarial learning strategy enables it to have the capability of extracting camera style-invariant pedestrian features, which effectively weakens the impact of camera style discrepancy on Baseline performance. On the two benchmark datasets, compared with the Baseline + CIEFM, Baseline + CIEFM + CSIL achieves improvements of 1.74% and 2.48% in Rank-1 accuracy, and 1.31% and 1.35% in mAP, respectively. Compared with the Baseline + CSIL, Baseline + CIEFM + CSIL achieves improvements of 2.17% and 1.46% in Rank-1 accuracy, and 1.27% and 1.63% in mAP, respectively. These results demonstrate that the combination of the camera style-invariant feature-level adversarial learning strategy and CIEFM can further improve the performance of VI-ReID models.

Table 4 Performance of different modules of CC-Net on the SYSU-MM01 and RegDB datasets

Baseline	CIEFM	CSIL	SYSU-MM01		RegDB	
			Rank-1	mAP	Rank-1	mAP
√			63.07	59.26	80.49	72.80
√	√		66.00	61.50	84.61	75.38
√		√	65.57	61.54	85.63	75.10
√	√	√	67.74	62.81	87.09	76.73

To assess the effectiveness of the fusion weight F', we conduct ablation experiments on the RegDB dataset, and the ablation experimental results are shown in Table 5. As can be seen from Table 5, the model that calculates the fusion weight F' in the way of formula (6) performs better than the model that calculates the fusion weight F' by directly adding F_{W1} and F_{W2}, with the Rank-1 accuracy improved by 1.65% and the mAP improved by 1.66%. This shows that calculating the fusion weight F' in the way of formula (6) is more effective.

To assess the impact of different loss combinations on the performance of our model, we conduct ablation experiments on the RegDB dataset and the experimental results are presented in Table 6. As illustrated in Table 6, compared

Table 5 Ablation experimental results of fusion weight F' calculation on the RegDB dataset

The computing of F'	Rank-1	mAP
$F' = F_{W1} + F_{W2}$	82.96%	73.72%
$F' = (F_{W1} \oplus F_{W2}) \otimes \sigma\left(\varphi_1\left(GAP\left(F_{W1} \oplus F_{W2}\right)\right) + \varphi_2\left(F_{W1} \oplus F_{W2}\right)\right)$	84.61%	75.38%

Table 6 Ablation experimental results of different loss combinations on the RegDB dataset

Model	Rank-1	mAP
Model with L_{ID}	54.81%	47.81%
Model with L_{ID} and L_{WRT}	80.49%	72.80%
Model with L_{ID}, L_{WRT} and $L_{cadv2}\left(W_C\right)$	83.30%	73.09%
Our model	87.09%	76.73%

with the model that only uses L_{ID}, the Rank-1 accuracy and mAP of our model are improved by 32.28% and 28.92%, respectively. Compared with the model with both L_{ID} and L_{WRT}, the Rank-1 accuracy and mAP of our model are improved by 6.6% and 3.93%, respectively. Compared with the model using L_{ID}, L_{WRT} and $L_{cadv2}\left(W_C\right)$, the Rank-1 accuracy and mAP of our model are improved by 3.79% and 3.64%, respectively. These fully prove that our camera style-invariant feature-level adversarial learning strategy is effective.

4.5 Visualization

4.5.1 Visualization of feature maps

We visualize the feature maps extracted by CC-Net and Baseline using Grad-CAM [71] on the SYSU-MM01 dataset, as shown in Fig. 4. In Fig. 4, the images in the first row are the original pedestrian images, the images in the second row are the feature maps of the corresponding original pedestrian images extracted by Baseline, and the images in the third row are the feature maps of the corresponding original pedestrian images extracted by CC-Net. From Fig. 4, it can be seen that Baseline focuses excessively on the local salient information of pedestrian images and ignores other effective information. For example, for the first visible pedestrian image, Baseline mainly focuses on the pedestrian's legs, and for the first infrared pedestrian image, Baseline mainly focuses on the pedestrian's feet. However, it can also be seen from Fig. 4 that CC-Net can focus on other effective information while focusing on the local salient information of pedestrian images. For example, for the first visible pedestrian image, CC-Net focuses not only on the pedestrian's legs but also on the pedestrian' arms. For the first infrared pedestrian image, CC-Net focuses not only on the pedestrian's feet, but also on the pedestrian's upper body.

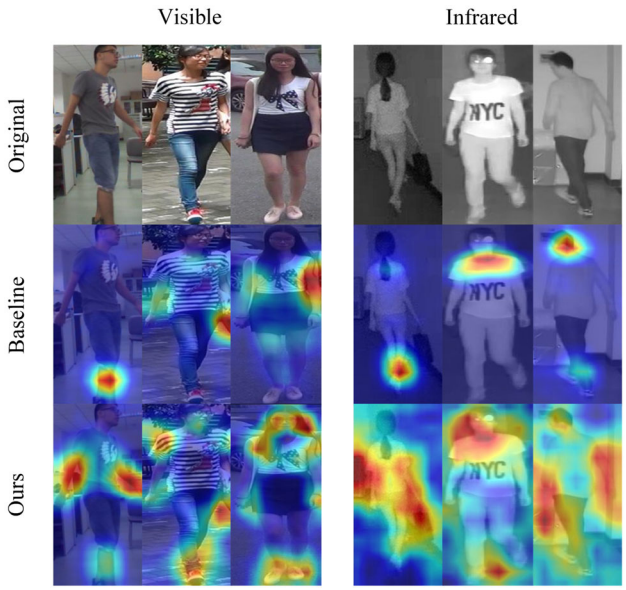

Fig. 4 Visualization of feature maps: the images in the first row are the original pedestrian images, the images in the second row are the feature maps of the corresponding original pedestrian images extracted by Baseline, and the images in the third row are the feature maps of the corresponding original pedestrian images extracted by CC-Net

4.5.2 Visualization of retrieval results

As shown in Fig. 5, we visualize part of the retrieval results of CC-Net, Baseline + CIEFM, Baseline + CSIL and Baseline on the RegDB dataset. Figure 5a shows the retrieval results in the visible-to-infrared mode, while Fig. 5b shows the retrieval results in the infrared-to-visible mode. In Fig. 5a, b, the first image in each row is the query image, and the rest are the 10 images retrieved from the gallery by Baseline, Baseline + CIEFM, Baseline + CSIL and CC-Net. In particular, the images marked with green boxes belong

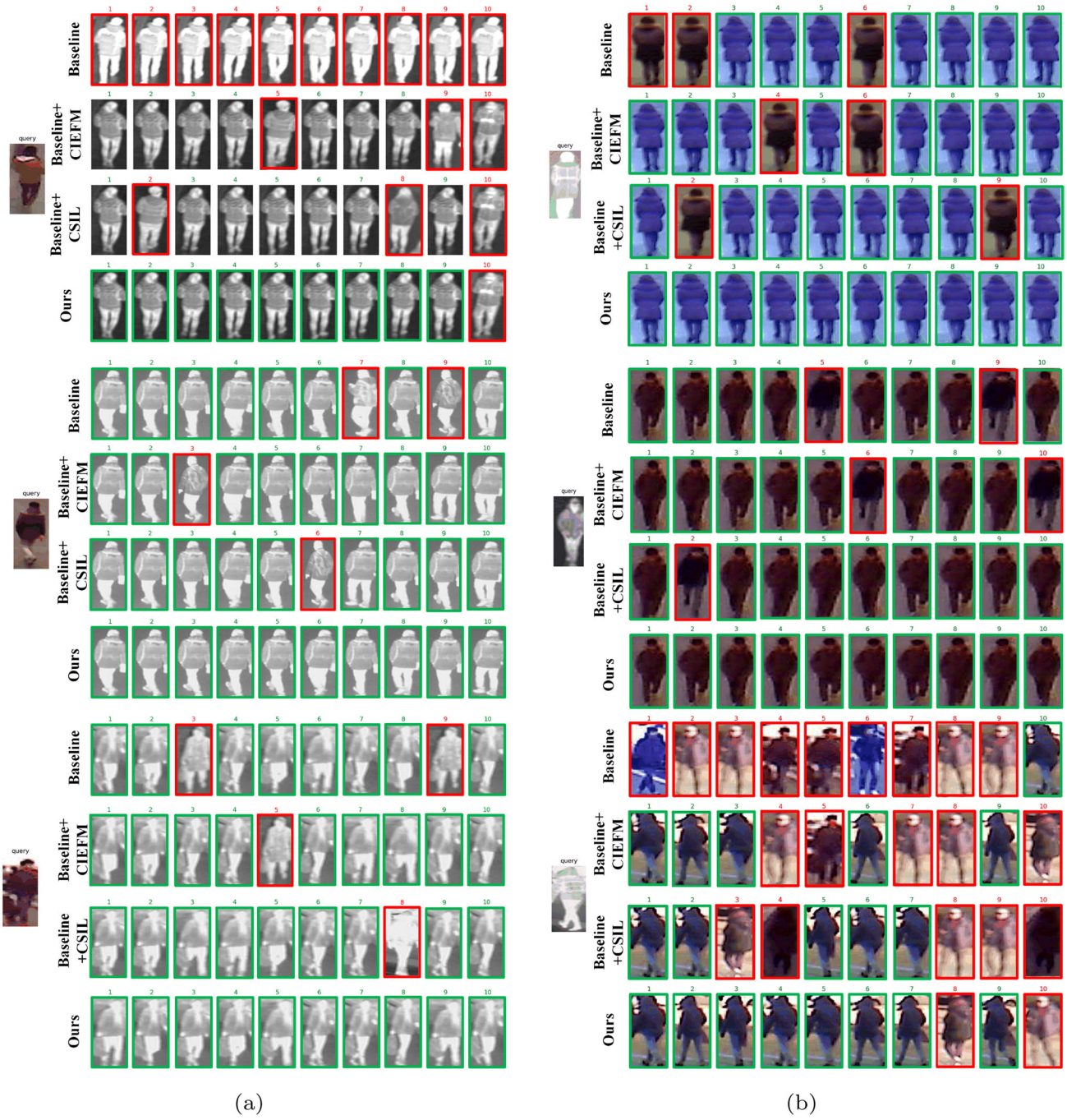

Fig. 5 Visualization of retrieval results: **a** the visible-to-infrared mode, and **b** the infrared-to-visible mode

to the same pedestrian as the corresponding query image, and the images marked with red boxes do not belong to the same pedestrian as the corresponding query image. Figure 5 shows that, compared with Baseline, CC-Net can significantly improve the ranking list. It can be also seen from Fig. 5, compared with Baseline, both Baseline + CIEFM and Baseline + CSIL effectively improve the accuracy of pedestrian retrieval, while CC-Net significantly improves the ranking list and basically achieves that the top 10 retrieved images belong to the same pedestrian as the corresponding query image. This demonstrates that CC-Net can still retrieve images that belong to the same pedestrian as the query image when the resolution of pedestrian images is low. However, for the first query image in Fig. 5a and the third query image in Fig. 5b, the images of other pedestrians also appear in the top 10 images retrieved by CC-Net, respectively. This may

be due to the limited ability of CC-Net to eliminate modality discrepancy caused by pixel-level image variations.

5 Conclusion

In this paper, we propose a novel camera style-invariant learning and channel interaction enhancement fusion network (CC-Net) for VI-ReID. By using the channel interaction enhancement fusion modules, the pedestrian features extracted by our network are more discriminative. Moreover, by training our network using the camera style-invariant feature-level adversarial learning strategy, our network has a certain capability of extracting camera style-invariant pedestrian features to weaken the impact of camera style discrepancy of pedestrian images on the performance of VI-ReID. Extensive experimental results on the SYSU-MM01 and RegDB datasets demonstrate that the performance of CC-Net achieves the recent advanced level. However, the ability of CC-Net to eliminate modality discrepancy caused by pixel-level image variations is limited, which may result in a poor performance in complex scenes with low foreground and background contrast. In our future work, we will investigate effective strategies to eliminate the impact of modality discrepancy on the performance of VI-ReID models.

Author Contributions HD did methodology, writing—original draft preparation, writing—reviewing and editing, supervision, funding acquisition. XH contributed to conceptualization, writing—reviewing and editing. YY gave software, visualization, and data curation. LH was involved in investigation, software, and validation. JG performed investigation, visualization, and validation. All authors have read and agreed to the published version of the manuscript.

Funding This work was supported in part by the Science and Technology Development Plan Project of Henan Province, China (No. 222102110135) and the Natural Science Foundation of Henan Province, China (No. 202300410093).

Data availability The data that support the findings of this study are available from the corresponding author upon reasonable request.

Code Availability The code that supports the findings of this study is available from the corresponding author upon reasonable request.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Zhu, X., Jing, X., You, X., Zuo, W., Shan, S., Zheng, W.: Image to video person re-identification by learning heterogeneous dictionary pair with feature projection matrix. IEEE Trans. Inf. Forensics Secur. **13**(3), 717–732 (2018)
2. Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., Hoi, S.: Deep learning for person re-identification: a survey and outlook. IEEE Trans. Pattern Anal. Mach. Intell. **44**(6), 2872–2893 (2022)
3. Li, Y., Jiang, X., Hwang, J.: Effective person re-identification by self-attention model guided feature learning. Knowl. Based Syst. **187**, 104832 (2020)
4. Bai, S., Tang, P., Torr, P., Latecki, L.: Re-ranking via metric fusion for object retrieval and person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 740–749 (2019)
5. Kalayeh, M., Basaran, E., Gokmen, M., Kamasak, M., Shah, M.: Human semantic parsing for person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1062–1071 (2018)
6. Wang, G., Yuan, Y., Chen, X., Li, J., Zhou, X.: Learning discriminative features with multiple granularities for person re-identification. In: Proceedings of the 26th ACM International Conference on Multimedia, pp. 274–282 (2018)
7. Zheng, F., Deng, C., Sun, X., Jiang, X., Guo, X., Yu, Z., Huang, F., Ji, R.: Pyramidal person re-identification via multi-loss dynamic training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8506–8514 (2019)
8. Hou, R., Ma, B., Chang, H., Gu, X., Shan, S., Chen, X.: Interaction-and-aggregation network for person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9309–9318 (2019)
9. Martinel, N., Foresti, G., Micheloni, C.: Deep pyramidal pooling with attention for person re-identification. IEEE Trans. Image Process. **29**, 7306–7316 (2020)
10. Luo, H., Gu, Y., Liao, X., Lai, S., Jiang, W.: Bag of tricks and a strong baseline for deep person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1487–1495 (2019)
11. Wu, A., Zheng, W., Yu, H., Gong, S., Lai, J.: Rgb-infrared cross-modality person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 5390–5399 (2017)
12. Liu, H., Ma, S., Xia, D., Li, S.: SFANet: a spectrum-aware feature augmentation network for visible-infrared person re-identification. IEEE Trans. Neural Netw. Syst. **34**(4), 1958–1971 (2023)
13. Liu, Q., He, X., Zhang, M., Teng, Q., Li, B., Qing, L.: Feature separation and double causal comparison loss for visible and infrared person re-identification. Knowl. Based Syst. **239**, 108042 (2022)
14. Cho, Y., Yoon, K.: Improving person re-identification via pose-aware multi-shot matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1354–1362 (2016)
15. Zhao, H., Tian, M., Sun, S., Shao, J., Yan, J., Yi, S., Wang, X., Tang, X.: Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 907–915 (2017)
16. Sarfraz, M., Schumann, A., Eberle, A., Stiefelhagen, R.: A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 420–429 (2018)
17. Cheng, D., Li, X., Qi, M., Liu, X., Chen, C., Niu, D.: Exploring cross-modality commonalities via dual-stream multi-branch network for infrared-visible person re-identification. IEEE Access. **8**, 12824–12834 (2020)
18. Guo, S., Xu, L., Feng, C., Xiong, H., Gao, Z., Zhang, H.: Multi-level semantic adaptation for few-shot segmentation on cardiac image sequences. Med. Image Anal. **73**, 102170 (2021)

19. Wu, Z., Allibert, G., Meriaudeau, F., Ma, C., Demonceaux, C.: Hidanet: Rgb-d salient object detection via hierarchical depth awareness. IEEE Trans. Image Process. **32**, 2160–2173 (2023)
20. Feng, J., Wu, A., Zheng, W.-S.: Shape-erased feature learning for visible-infrared person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 22752–22761 (2023)
21. Lan, L., Teng, X., Zhang, J., Zhang, X., Tao, D.: Learning to purification for unsupervised person re-identification. IEEE Trans. Image Process. **32**, 3338–3353 (2023)
22. Zhu, Y., Yang, Z., Wang, L., Zhao, S., Hu, X., Tao, D.: Hetero-center loss for cross-modality person re-identification. Neurocomputing **386**, 97–109 (2020)
23. Sun, J., Li, Y., Chen, H., Peng, Y., Zhu, X., Zhu, J.: Visible-infrared cross-modality person re-identification based on whole-individual training. Neurocomputing **440**, 1–11 (2021)
24. Ran, L., Hong, Y., Zhang, S., Yang, Y., Zhang, Y.: Improving visible-thermal ReID with structural common space embedding and part models. Pattern Recogn. Lett. **142**, 25–31 (2021)
25. Zhang, J., Li, X., Chen, C., Qi, M., Wu, J., Jiang, J.: Global-local graph convolutional network for cross-modality person re-identification. Neurocomputing **452**, 137–146 (2021)
26. Liu, H., Cheng, J., Wang, W., Su, Y., Bai, H.: Enhancing the discriminative feature learning for visible-thermal cross-modality person re-identification. Neurocomputing **398**, 11–19 (2020)
27. Xiang, S., Chen, H., Ran, W., Yu, Z., Liu, T., Qian, D., Fu, Y.: Deep Multimodal Fusion for Generalizable Person Re-identification **18**(9), 1–9 arXiv:2211.00933 (2022)
28. Ye, M., Wang, Z., Lan, X., Yuen, P.: Visible thermal person re-identification via dual-constrained top-ranking. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI), pp. 1092–1099 (2018)
29. Ye, M., Lan, X., Li, J., Yuen, P.: Hierarchical discriminative learning for visible thermal person re-identification. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence(AAAI), pp. 7501–7508 (2018)
30. Feng, Z., Lai, J., Xie, X.: Learning modality-specific representations for visible-infrared person re-identification. IEEE Trans. Image Process. **29**, 579–590 (2020)
31. Dai, P., Ji, R., Wang, H., Wu, Q., Huang, Y.: Cross-modality person re-identification with generative adversarial training. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI), pp. 677–683 (2018)
32. Xia, D., Liu, H., Xu, L., Wang, L.: Visible-infrared person re-identification with data augmentation via cycle-consistent adversarial network. Neurocomputing **443**, 35–46 (2021)
33. Wang, Z., Wang, Z., Zheng, Y., Chuang, Y., Satoh, S.: Learning to reduce dual-level discrepancy for infrared-visible person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 618–626 (2019)
34. Liao, S., Hu, Y., Zhu, X., Li, S.: Person re-identification by Local Maximal Occurrence representation and metric learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 07-12-June, pp. 2197–2206 (2015)
35. Ke, Q., Bennamoun, M., Rahmani, H., An, S., Sohel, F., Boussaid, F.: Identity adaptation for person re-identification. IEEE Access. **6**, 48147–48155 (2018)
36. Lin, Y., Zheng, L., Zheng, Z., Wu, Y., Hu, Z., Yan, C., Yang, Y.: Improving person re-identification by attribute and identity learning. Pattern Recogn. **95**, 151–161 (2019)
37. Cheng, D., Gong, Y., Zhou, S., Wang, J., Zheng, N.: Person re-identification by multi-channel parts-based CNN with improved triplet loss function. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1335–1344 (2016)
38. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 (2017)
39. Liao, S., Li, S.: Efficient psd constrained asymmetric metric learning for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 3685–3693 (2015)
40. Liu, J., Sun, Y., Zhu, F., Pei, H., Yang, Y., Li, W.: Learning Memory-Augmented Unidirectional Metrics for Cross-modality Person Re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 19344–19353 (2022)
41. Li, X., Lu, Y., Liu, B., Liu, Y., Yin, G., Chu, Q., Huang, J., Zhu, F., Zhao, R., Yu, N.: Counterfactual intervention feature transfer for visible-infrared person re-identification. In: Proceedings of the European Conference on Computer Vision, pp. 381–398. Springer (2022)
42. Huang, N., Liu, J., Miao, Y., Zhang, Q., Han, J.: Deep learning for visible-infrared cross-modality person re-identification: a comprehensive review. Information Fusion. **91**, 396–411 (2023)
43. Ye, M., Shen, J., Crandall, D., Shao, L., Luo, J.: Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In: Lecture Notes in Computer Science (including Subbooktitle Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pp. 299–247 (2020)
44. Ye, M., Lan, X., Leng, Q., Shen, J.: Cross-modality person re-identification via modality-aware collaborative ensemble learning. IEEE Trans. Image Process. **29**, 9387–9399 (2020)
45. Zhang, Y., Yan, Y., Lu, Y., Wang, H.: Towards a unified middle modality learning for visible-infrared person re-identification. In: Proceedings of the 29th ACM International Conference on Multimedia, pp. 788–796 (2021)
46. Liu, H., Tan, X., Zhou, X.: Parameter sharing exploration and hetero-center triplet loss for visible-thermal person re-identification. Knowl. Based Syst. IEEE Trans. Multimedia. **23**, 4414–4425 (2021)
47. Wu, Q., Dai, P., Chen, J., Lin, C., Wu, Y., Huang, F., Zhong, B., Ji, R.: Discover Cross-Modality Nuances for Visible-Infrared Person Re-Identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4328–4337 (2021)
48. Choi, S., Lee, S., Kim, Y., Kim, T., Kim, C.: Hi-cmd: Hierarchical cross-modality disentanglement for visible-infrared person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10257–10266 (2020)
49. Zhang, S., Yang, Y., Wang, P., Liang, G., Zhang, X., Zhang, Y.: Attend to the difference: cross-modality person re-identification via contrastive correlation. IEEE Trans. Image Process. **30**, 8861–8872 (2021)
50. Wan, L., Sun, Z., Jing, Q., Chen, Y., Lu, L., Li, Z.: G2da: geometry-guided dual-alignment learning for RGB-infrared person re-identification. Pattern Recogn. **135**, 109150 (2023)
51. Sun, Z., Zhao, F.: Counterfactual attention alignment for visible-infrared cross-modality person re-identification. Pattern Recogn. Lett. **168**, 79–85 (2023)
52. Wang, G., Zhang, T., Cheng, J., Liu, S., Yang, Y., Hou, Z.: Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 3622–3631 (2019)
53. Wang, G., Yang, Y., Zhang, T., Cheng, J., Hou, Z., Tiwari, P., Pandey, H.: Cross-modality paired-images generation and augmentation for RGB-infrared person re-identification. Neural Netw. **128**, 294–304 (2020)
54. Wu, H., Zhang, J., Huang, K., Liang, K., Yu, Y.: FastFCN: Rethinking Dilated Convolution in the Backbone for Semantic Segmentation. arXiv preprint arXiv:1903.11816 (2019)

55. Xiong, L., Karlekar, J., Zhao, J., Cheng, Y., Xu, Y., Feng, J., Pranata, S., Shen, S.: A good practice towards top performance of face recognition: transferred deep feature fusion. arXiv preprint arXiv:1704.00438 (2017)
56. Nguyen, D., Hong, H., Kim, K., Park, K.: Person recognition system based on a combination of body images from visible light and thermal cameras. Sensors **17**(3), 605 (2017)
57. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016)
58. Deng, J., Dong, W., Socher, R., Li, L., Li, K., Li, F.: Imagenet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255 (2009)
59. Wei, Z., Yang, X., Wang, N., Gao, X.: Syncretic modality collaborative learning for visible infrared person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 225–234 (2021)
60. Ye, M., Lan, X., Leng, Q.: Modality-aware collaborative learning for visible thermal person re-identification. In: Proceedings of the 27th ACM International Conference on Multimedia, pp. 347–355 (2019)
61. Ye, M., Lan, X., Wang, Z., Yuen, P.: Bi-directional center-constrained top-ranking for visible thermal person re-identification. IEEE Trans. Inf. Forensics Secur. **15**, 407–419 (2020)
62. Li, D., Wei, X., Hong, X., Gong, Y.: Infrared-visible cross-modal person re-identification with an x modality. In: Proceedings of the AAAI Conference on Artificial Intelligence(AAAI), pp. 4610–4617 (2020)
63. Wang, G., Zhang, T., Yang, Y., Cheng, J., Chang, J., Liang, X., Hou, Z.: Cross-modality paired-images generation for RGB-infrared person re-identification. In: Proceedings of the Association for the Advance of Artificial Intelligence (AAAI), pp. 12144–12151 (2020)
64. Basaran, E., Gokmen, M., Kamasak, M.: An efficient framework for visible-infrared cross modality person re-identification. Signal Process. Image Commun. **87**, 115933 (2020)
65. Zhang, C., Liu, H., Guo, W., Ye, M.: Multi-scale cascading network with compact feature learning for rgb-infrared person re-identification. In: Proceedings of the 25th International Conference on Pattern Recognition (ICPR), pp. 8679–8686 (2021)
66. Cai, X., Liu, L., Zhu, L., Zhang, H.: Dual-modality hard mining triplet-center loss for visible infrared person re-identification. Knowl. Based Syst. **215**, 106772 (2021)
67. Ye, M., Shen, J., Shao, L.: Visible-infrared person re-identification via homogeneous augmented tri-modal learning. IEEE Trans. Inf. Forensics Secur. **16**, 728–739 (2021)
68. Lu, Y., Wu, Y., Liu, B., Zhang, T., Li, B., Chu, Q., Yu, N.: Cross-Modality Person Re-Identification With Shared-Specific Feature Transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 13376–13386 (2020)
69. Hao, X., Zhao, S., Ye, M., Shen, J.: Cross-Modality Person Re-Identification via Modality Confusion and Center Aggregation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 16383–16392 (2021)
70. Chen, Y., Wan, L., Li, Z., Jing, Q., Sun, Z.: Neural feature search for rgb-infrared person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 587–597 (2021)
71. Selvaraju, R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: visual explanations from deep networks via gradient-based localization. Int. J. Comput. Vision **128**(2), 336–359 (2020)