SCFNet: A Spatial-Channel Features Network Based on Heterocentric Sample Loss for Visible-Infrared Person Re-Identification

Peng Su, Rui Liu(✉), Jing Dong, Pengfei Yi, and Dongsheng Zhou

Key Laboratory of Advanced Design and Intelligent Computing Ministry of
Education, School of Software Engineering,
Dalian University, Dalian, China
liurui@dlu.edu.cn

Abstract. Cross-modality person re-identification between visible and infrared images has become a research hotspot in the image retrieval field due to its potential application scenarios. Existing research usually designs loss functions around samples or sample centers, mainly focusing on reducing cross-modality discrepancy and intra-modality variations. However, the sample-based loss function is susceptible to outliers, and the center-based loss function is not compact enough between features. To address the above issues, we propose a novel loss function called Heterocentric Sample Loss. It optimizes both the sample features and the center of the sample features in the batch. In addition, we also propose a network structure combining spatial and channel features and a random channel enhancement method, which improves feature discrimination and robustness to color changes. Finally, we conduct extensive experiments on the SYSU-MM01 and RegDB datasets to demonstrate the superiority of the proposed method.

1 Introduction

Person re-identification(ReID) is a popular research direction in the image retrieval field, which is a cross-device pedestrian retrieval technology. Person ReID faces many challenges due to factors such as lighting changes, viewpoint changes, and pose changes. Most existing methods [1,2,3,4] focus on matching person images with visible images. However, when at night, it can be hard for visible cameras to capture clear pictures of pedestrians due to insufficient light. Nowadays, many new outdoor monitors have integrated infrared image capture devices. Therefore, how to use both visible and infrared images for cross-modality person re-identification (VI-ReID) has become a very significant research issue.

Compared with the single-modality person ReID task, VI-ReID is more challenging with a higher variation in data distribution. Two classic methods have been investigated to solve the VI-ReID problem. The first method is based on modal conversion [5,6], which eliminates modal discrepancy by converting the images of two modalities to each other. However, the operation of modal conversion is relatively complicated. And inevitably, some key information is lost and some noise is introduced in the conversion process. Another method is based on representation learning [7,8]. This method often maps visible and infrared images into a unified feature space and then learns a discriminative feature representation for each person. Among them, local features as a common feature representation method have been widely used in person ReID. Inspired by previous work MPANet [9], we propose a Spatial-Channel Features Network (SCFNet) structure based on local feature representation in this paper. SCFNet extracts local features in simultaneously spatial and channel dimensions, enhancing the feature representation capability. Besides the above two methods, metric learning [10,11] is often used in VI-ReID as a key technique. In particular, triplet loss and its variants [10,12] are the most dominant metric learning methods. However, most of the existing triplet losses are designed around samples or sample centers. The loss designed around the sample is vulnerable to the influence of anomalous samples, and the loss designed around the center is not tight enough. To solve this problem, an improved heterocentric sample loss is proposed in this paper. It can tolerate the outliers between samples and also takes into account the modal differences. More importantly, it can make the distribution of sample features relatively more compact.

In addition, we also propose a local random channel enhancement method for VI-ReID. By randomly replacing pixel values in local areas of a color image, the model is made more robust to color changes. We combine it with global random grayscale to improve the performance of the model with only a small increase in computational effort.

The summary of our main contributions is as follows:

- A SCFNet network structure is proposed that allows local features to be extracted from the spatial and channel levels. In addition, a local random channel enhancement is used to further increase the identification and robustness of the features.
- A heterocentric sample loss is proposed, which optimizes the structure of the feature space from both sample and center perspectives, enhancing intra-class tightness and inter-class separability.
- Experiments on two benchmark datasets show that our proposed method obtain good performance in the VI-ReID tasks. In particular, 69.37% of mAP and 72.96% of Rank1 accuracy are obtained on SYSU-MM01.

2 Related Work

VI-ReID was first proposed by WU et al. [13] in 2017. They contributed a standard dataset SYSU-MM01 and designed a set of evaluation criteria for this problem. Recently, the research on VI-ReID has been divided into two main types: modality conversion and representation learning. In addition, metric learning, as a key algorithm used in these two methods, is often studied separately. This paragraph will introduce the related work from these three aspects.

Modality conversion based methods. This method usually interconverts images of two modalities by GAN to reduce the modal difference. Wang et al.

[5] proposed AlignGAN to transform visible images into infrared images, and they used pixel and feature alignment to alleviate intra- and inter-modal variation. In the second year, they also used GAN to generate cross-modal paired images in [14] to solve the VI-ReID problem by performing set-level and instance-level feature alignment. Choi et al. [15] used GAN to generate cross-modality images with different morphologies to learn decoupled representations and automatically decouple identity discriminators and identity irrelevant factors from infrared images. Hao et al. [16] designed a module to distinguish images for better performance in cross-modality matching using the idea that generators and discriminators are mutually constrained in GAN. The method based on modality conversion reduces inter-modal differences to a certain extent, but inevitably it also introduces some noise. In addition, methods based on modality conversion are relatively complex and usually require more training time.

Representation learning based methods. The method of representation learning aims to use a feature extraction structure to map the features of both modalities into the unified feature space. Ye et al. [7] proposed an AGW baseline network, which used Resnet50 as a feature extraction network and added a non-local attention module to make the model focus more on global information. In the same year, they also proposed a DDAG [8] network to extract contextual features by intra-modal weighted aggregation and inter-modal attention. Liu et al. [17] incorporated feature skip connections in the middle layers into the latter layers to improve the discrimination and robustness of the person feature. Huang et al. [18] captured the modality-invariant relationship between different character parts based on appearance features, as a supplement to the modality-shared appearance features. Zhang et al. [19] proposed a multi-scale part-aware cascade framework that aggregates local and global features of multiple granularities in a cascaded manner to enrich and enhance the semantic information of features. Representation learning approaches often achieve good results through well-designed feature extraction structures. And, the representation of person features has evolved from the initial global features to richer features, such as local features, multi-scale features, etc.

Metric Learning. Metric learning aims to learn the degree of similarity between person images. As a key algorithm in VI-ReID, metric learning has been used in both modal transformation methods and representation learning methods. In VI-ReID, triplet loss and its variants are the most used metric learning methods. Ye et al. [10] proposed the BDTR loss, which dealt with both cross-modal variation and intra-modal variation to ensure feature discriminability. Li et al. [12] proposed a strategy of batch sampling all triples for the imbalance problem of modal optimization in the optimization process of triplet loss. Liu et al. [20] proposed Heterocentric Triplet Loss (HcTri) to improve intra-modality similarity by constraining the center-to-center distance of two heterogeneous intra-class modalities. Wu et al. [9] introduced a Center Cluster Loss (CC) to study the relationship between identities. The CC loss establishes the relationship between class centers on the one hand, and clusters the samples toward the center on the other hand. Inspired by the work of [9] and [20], we design a heterocentric sample loss that combines samples and sample centers. This loss makes the sample distribution more compact while optimizing the intra-class similarity and inter-class difference, which further enhances the identification and robustness of features.

3 Methodology

3.1 Network Architecture

We chose the MPANet as our baseline model. It has 3 main characteristics, Modality Alleviation Module (MAM) eliminates the modality-related information, Pattern Alignment Module (PAM) discovers nuances in images to increase the recognizability of features, and Modal Learning (ML) mitigates modal differences with mutual mean learning. We chose this baseline for two main reasons. First, it provides a framework to learn modal invariant features of the person. Secondly, the performance of this network is very superior.

The architecture of SCFNet is shown in Fig. 1. First, we preprocess the visible images using local random channel enhancement(CA) and random grayscale(RG). Then the infrared image and the processed visible image are put into a backbone network consisting of Resnet50 and MAM to extract features. The 3D features map obtained are represented $X \in \mathbb{R}^{C \times H \times W}$. In MPANet, the authors use the PAM to extract nuances features from multiple patterns in the channel dimension. But in the spatial dimension, some detailed features of a person should also be included. Therefore, we introduce a spatial feature extraction module to extract local information about the person from the space dimension.

To extract the spatial feature, a simple method referring to [20] and [21] is to cut the feature map horizontally and extract each small piece of feature as

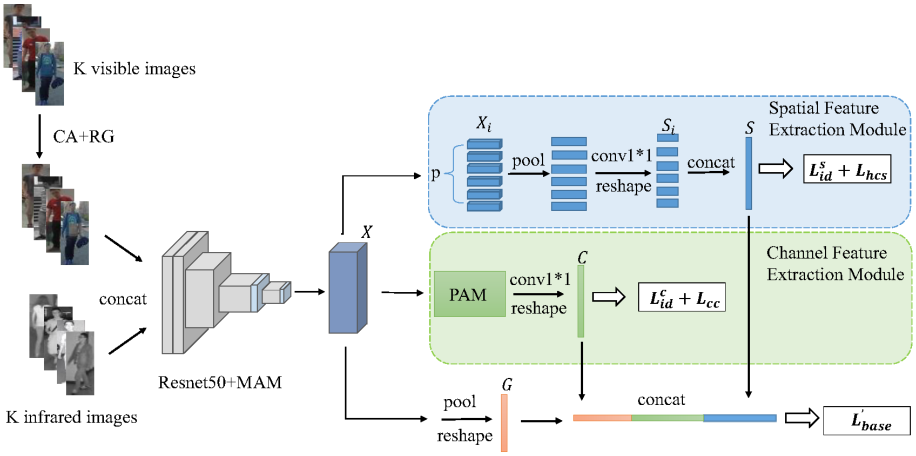


Fig. 1. The architecture of the proposed SCFNet.

a representation of local features. As shown in Fig. 1, we divide the obtained 3D feature map X in the horizontal direction, evenly into p parts. The features of each small segment are denoted as $X_i \in \mathbb{R}^{C\times\frac{H}{p}\times W}, i \in \{1, 2, \ldots, p\}$. Then, each local feature is reshaped into a d-dimensional vector using pooling and convolution operations. The formula is expressed as $S_i \in \mathbb{R}^d, i \in \{1, 2, \ldots, p\}$, where

$$S_i = Reshape(Conv(pool(X_i))) \tag{1}$$

Finally, all local features are concatenated as the final spatial feature representation $S \in \mathbb{R}^{D_1}$, where $D_1 = p \times d$.

For channel features, we directly use the PAM in MPANet. However, to decrease the number of parameters during computation, we downsampled the features by adding a 1×1 convolutional layer at the end. Finally, the channel features obtained after the reshaping operation are represented as $C \in \mathbb{R}^{D_2}$. For the convenience of calculation, the value of D_2 we set is equal to D_1.

Since local features are susceptible to factors like occlusion and pose variations, we also incorporate global features. The global feature representation $G \in \mathbb{R}^{D_3}$ is obtained by directly using pooling and reshaping operations on the 3D feature X. Finally, we join S, C, and G as the final feature representation in the inference stage.

3.2 Loss Function

Triplet loss serves an essential role in VI-ReID. Traditional triplet losses are designed by constraining the samples, but the heterocentric triplet loss [20] replaces the samples with the center of each class. It is formulated as follows:

$$\begin{aligned} L_{hc\_tri} = &\sum_{i=1}^{P}[\rho + ||c_v^i - c_t^i||_2 - \min_{\substack{n\in\{v,t\}\\ j\neq i}} ||c_v^i - c_n^j||_2]_+ \\ &+ \sum_{i=1}^{P}[\rho + ||c_t^i - c_v^i||_2 - \min_{\substack{n\in\{v,t\}\\ j\neq i}} ||c_t^i - c_n^j||_2]_+ \end{aligned} \tag{2}$$

where ρ is the margin parameter for heterocentric triplet loss, $[x]_+ = max(x, 0)$ denotes the standard hinge loss, and $||x_a - x_b||_2$ denotes the Euclidean distance of two feature vectors x_a and x_b. P denotes the total number of different classes in a mini-batch. c_v^i and c_t^i denote the central features of the visible and infrared class i in a mini-batch, respectively. They are obtained by averaging the samples of all classes i in the respective modalities, which are calculated as follows:

$$c_v^i = \frac{1}{K}\sum_{j=1}^{K} v_j^i \tag{3}$$

$$c_t^i = \frac{1}{K}\sum_{j=1}^{K} t_j^i \tag{4}$$

where v_j^i and t_j^i denote the feature representation of the jth visible image and the jth infrared image of class i respectively.

The heterocentric triplet loss uses class centers instead of samples, relaxing the tight constraint in traditional triplet loss and alleviating the effects caused by anomalous samples, allowing the network to converge better. However, optimizing only the central features means that the change for each sample feature is small. If the mini-batch is large, then the constraints assigned to each sample are small. This will cause the network to have a slow convergence rate. Therefore, we propose to optimize the central features while also applying constraints to each sample, so that each sample is closer to its central features. We refer to this as heterocentric sample loss, which calculation formula is shown in Eq. 5. where δ is a balance factor.

$$\begin{aligned}
L_{hc\_tri} = & \sum_{i=1}^{P} [\rho + ||c_v^i - c_t^i||_2 - \min_{\substack{n\in\{v,t\}\\ j\neq i}} ||c_v^i - c_n^j||_2]_+ \\
& + \sum_{i=1}^{P} [\rho + ||c_t^i - c_v^i||_2 - \min_{\substack{n\in\{v,t\}\\ j\neq i}} ||c_t^i - c_n^j||_2]_+ \\
& + \delta \sum_{j=1}^{K} [||v_j^i - c_v^i||_2 + ||t_j^i - c_t^i||_2] \qquad (5)
\end{aligned}$$

The heterocentric sample loss is based on the heterocentric triplet loss and adds the constraint that the Euclidean distance between the sample feature and the central feature should be as close as possible. In this way, the samples of the same class will be more compact in the feature space. The difference between the heterocentric triplet loss and the heterocentric sample loss is shown in Fig. 2(a) and (b).

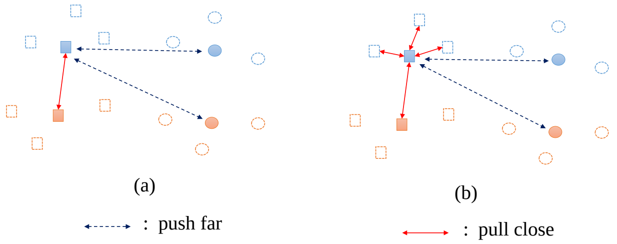


Fig. 2. Comparison of Heterocentric Triplet Loss (a) and Heterocentric Sample Loss (b). Different colors represent different modalities, and different shapes represent different classes. Graphs with dashed borders and no color fills represent sample characteristics. A color-filled graph represents the central feature computed from the sample features of the corresponding color.

Heterocentric sample loss reduces both cross-modality variation and intermodal variation in a simple way. It has two main advantages: (1) It is not a strong constraint, so it can be robust to abnormal samples. (2) By constraining the relationship between samples and centers, each sample is brought near the center of its class, thus making the final features more compact and discriminative, and also speeding up the convergence rate.

Besides triplet loss, identity loss is also often used in the VI-ReID. Given an image y, p_i denotes the class predicted by the network, q_i denotes the true label, and the identity loss is denoted as:

$$L_{id} = -\sum_{1}^{n} q_i \log p_i \tag{6}$$

Let the total loss of the baseline model MPANet be denoted as L_{base}, from which the central cluster loss is removed and denoted as $L_{base}^{'}$. We use the identity loss L_{id}^{s} and heterocentric sample loss L_{hcs} for spatial features, the central cluster loss L_{cc} and identity loss L_{id}^{c} for channel features, and $L_{bass}^{'}$ for total features (as shown in Fig. 1). Finally, the total loss of SCFNet is defined as:

$$Loss = L_{bass}^{'} + L_{hcs} + L_{id}^{s} + L_{cc} + L_{id}^{c} \tag{7}$$

3.3 Local Random Channel Enhancement

There is naturally a large difference in appearance between infrared and visible images, which is the most significant cause of modal differences. We observe that the infrared images and the images extracted from each channel of visible images separately are similar to the grayscale images in appearance. Therefore, we propose a local random channel enhancement method to alleviate the influence of modal differences.

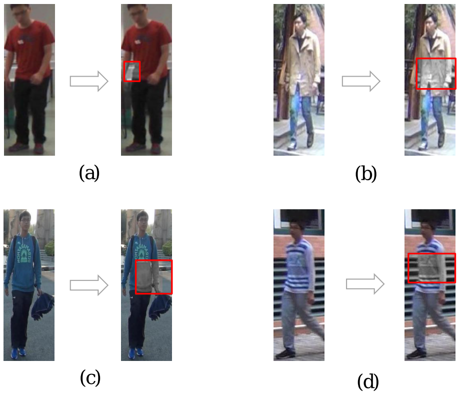


Fig. 3. Local Random channel enhancement. a, b, c, and d correspond to the 4 conversions, respectively.

The specific operation of local random channel enhancement is as follows. Given a visible image, we randomly select a region in the visible light image and randomly replace the pixel value of the region with one of the pixel values of the R, G, B channel or the gray value. Fig. 3 shows the effect before and after the transformation.

Local random channel enhancement makes visible and infrared images more similar in appearance by changing the pixel values of the images, while this random enhancement also forces the network to be less sensitive to color changes and facilitates its learning of color-independent features. In our experiments, we use a combination of random channel enhancement and random grayscale. With only a slight increase in computational effort, the model's performance can be effectively improved.

4 Experiments

4.1 Experimental Setting

Datasets. We conducted experiments on two public VI-ReID datasets, SYSU-MM01 [5] and RegDB [22]. For a fair comparison, we report the results of Rank-k and mAP on each dataset.

SYSU-MM01 is the first and the largest public dataset proposed for VI-ReID. It was captured by 6 cameras. The training set has 22258 visible pictures and 11909 infrared pictures and the testing set has 3803 infrared pictures and 301 visible pictures. We followed the evaluation protocol proposed in [5] for the SYSU-MM01 and mainly reported the results for single and multi-shot settings in all and indoor search modes.

The RegDB dataset has 412 person IDs with 10 visible and 10 infrared images for each person. 206 person IDs are randomly selected for training in the training phase and the remaining 206 person IDs are used for testing. For the RegDB dataset, we used the two most common search settings: "Infrared to Visible" and "Visible to Infrared".

Implementation details. This paper implements an improved model based on Pytorch. All evaluation indicators are trained on the Sitonholy SCM artificial intelligence cloud platform, using a single Tesla P100-SXM2 graphics card on Ubuntu operating system. Our network uses the Adam optimizer with training epochs of 140 generations and an initial learning rate of 3.5×10^{-4}, at the 80th and 120th generations time decay. In the spatial feature extraction module, the number of blocks that the feature map is divided is 6, and each local feature is represented as a 512-dimensional vector. Therefore, the final dimension of the spatial feature is $D_1 = 3072$. In the channel feature extraction module, the dimension D_2 of the downsampled channel feature is also set to 3072. The final dimension $D_3 = 2048$ of the global feature. For the loss of heterocentric samples, after experimental comparison, we set the marginal ρ to 0.9 and δ to 0.1 on the SYSU-MM01 dataset and set the marginal ρ to 0.3 and δ to 1 on the RegDB dataset. The probabilities used for local random channel enhancement and global grayscale are 0.8 and 0.3, respectively.

4.2 Ablation Experiment

Effectiveness of the components. To validate the effectiveness of each module of SCFNet, we conducted ablation experiments on two public datasets. The SYSU-MM01 dataset tests the single-shot settings in all search mode, and the RegDB dataset tests the settings from visible to infrared. The results of the ablation experiments are shown in Table 1 and Table 2. Among them, "CL" indicates that the channel feature extraction module is used, "SL" indicates that the spatial feature extraction module is used, "CA" indicates that the local random channel enhancement is used, and "HCS" indicates that the heterocentric sample loss is used. For the RegDB dataset, the spatial feature extraction module is not used and the Center Cluster Loss of the channel feature extraction module is replaced by the heterocentric sample loss. This is because RegDB is a small dataset, using the channel feature extraction module is enough to extract effective features, and using a too complex structure will increase the difficulty of network convergence.

From the comparison between version 1 and version 2 in Table 1, it can be found that using the spatial feature extraction module can improve Rank1 and mAP by 1.22 and 0.88, respectively, compared to the situation of not using it. This illustrates the effectiveness of the method to mine person feature information from the spatial dimension. From version 2 and version 3 in Table 1, and version 1 and version 2 in Table 2, it can be seen that using local random channel enhancement can enhance the model performance. This indicates that the use of local random channel enhancement can reduce the sensitivity of the model to color. From the comparison of version 2 and version 4 in Table 1, and version 1 and version 3 in Table 2, it can be seen that the use of heterocentric sample

Table 1. Performance evaluation of each component on the SYSU-MM01 dataset.

Version	CL	SL	CA	HCS	Rank1	mAP
Version1	✓				68.14	65.17
Version2	✓	✓			69.36	66.05
Version3	✓	✓	✓		71.46	67.36
Version4	✓	✓		✓	71.60	67.80
Version5	✓	✓	✓	✓	72.96	69.37

Table 2. Performance evaluation of each component on the RegDB dataset.

Version	CL	CA	HCS	Rank1	mAP
Version1	✓			83.11	80.58
Version2	✓	✓		84.91	81.61
Version3	✓		✓	84.52	81.11
Version4	✓	✓	✓	86.33	82.10

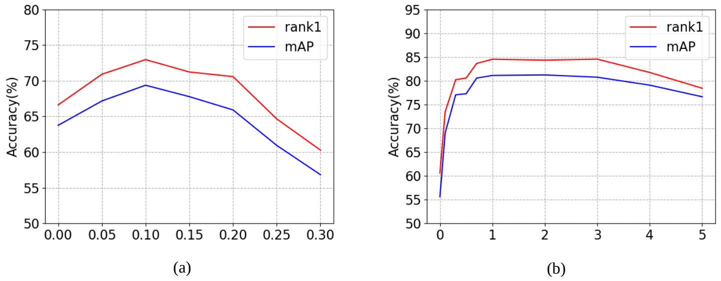


Fig. 4. Evaluating the weight parameter δ in the loss of heterocentric samples. (a) shows the results in SYSU-MM01, and (b) shows the results in RegDB.

loss can improve the mAP by 1.75 and the Rank1 by 2.24 on SYSU-MM01 and improve the mAP by 0.53 and the Rank1 by 1.41 on RegDB. This is because the heterocentric sample loss reduces intra-class separability and makes the model feature distribution more compact. So it is easier to distinguish different classes. We can see from Table 1 version 5 and Table 2 version 4, the best results were achieved by using these modules together. This illustrates that these modules are complementary in terms of performance and combining them allows us to achieve the highest performance of our model.

Discussion on δ of Heterocentric Sample Loss. The hyperparameter δ is a balancing factor whose value can greatly affect the performance of the model. So we conduct several experiments to determine the value of δ in Eq 5. The experimental results are shown in Fig 4.

On the SYSU-MM01 dataset, the network is very sensitive to changes in δ. We changed δ from 0 to 0.3, as can be seen from Fig 4(a), when δ is less than 0.1, the Rank1 and mAP of the network tend to rise, and when δ is greater than 0.1, the performance of the network begins to gradually decline, especially when $\delta = 0.3$, its Rank1 and mAP drop to around 60% and 55%, respectively. Therefore, on this dataset, δ is taken as 0.1. On the RegDB dataset, the change of the network to δ is quite different from SYSU-MM01. When δ takes 0, the loss function degenerates into a heterocentric loss, and in our network structure, Rank1 and mAP are only about 61% and 56%. When δ takes a small value of 0.1, the performance of the network increases rapidly, which also shows the effectiveness of the loss of heterocentric samples. When the value of δ is between 0 and 1, the performance of the network gradually increases. When the value of δ is between 1 and 3, the Rank1 and mAP of the network only change slightly. When the value of δ is greater than 3, the Rank1 and mAP of the network begin to have a downward trend. So, on this dataset, we set the value of δ to 1.

For this phenomenon, we speculate that this is due to the different data distributions in these two datasets. In the RegDB dataset, there is a one-to-one correspondence between the poses in the infrared image and the visible image

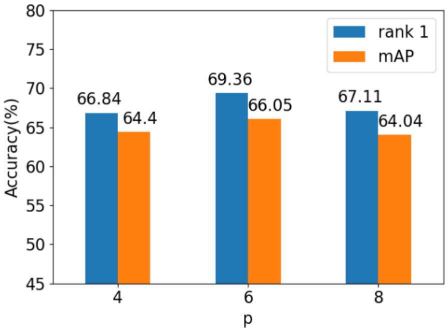
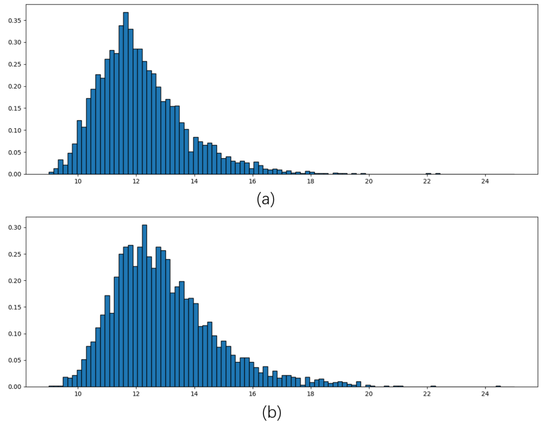


Fig. 5. Influence of the number p of spatially local features



(a)

(b)

Fig. 6. The histogram of sample-to-center distance.

for the same person, so its intra-class discrepancy is smaller than that of SYSU-MM01. The δ control item in Eq 5 is to shorten the distance between the sample and the center and reduce the intra-class discrepancy. Therefore, the RegDB dataset is less sensitive to δ changes than the SYSU-MM01 dataset.

Discussion on the number p of spatial feature divisions. The number of spatial features that are divided determines the size of each local feature and also affects the representability of each local feature. We design some experiments to compare the number of segmented spatial features. Experiments are carried out with the number of divisions being 4, 6, and 8 respectively in SYSU-MM01 dataset. The experimental results are shown in Fig 5. It can be seen that when the number of divisions is 6, the effect is the best, with Rank1 and mAP being 69.36% and 66.05%, respectively. When the number of divisions is 4, the granularity of the obtained local features is not fine enough, and it is difficult to capture more detailed features. The experimental results drop again when the number of divisions is 8. This is because features with too small a granularity are easily affected by noise, so it is hard for the network to capture effective local information. After comparison, the number of divisions of 6 is a reasonable value, so in the experiment, we take p=6.

Visualization of sample-to-center distance. To demonstrate the performance of heterocentric sample loss more precisely, we train two models on the spatial branch of SCFNet using heterocentric triplet loss and heterocentric sample loss respectively. Then, the data from all test sets are fed into the network and the distance to the center is calculated for each category. These distances

Table 3. Comparison of SYSU-MM01 dataset with state-of-the-art methods in all search and indoor search modes. 1^{st} and 2^{nd} best results are marked in red and blue, respectively.

Method	Single-Shot				Multi-Shot			
	R1	R10	R20	mAP	R1	R10	R20	mAP
All Search								
Zero-Pad[13]	14.80	54.12	71.33	15.95	19.13	61.4	78.41	10.89
cmGAN[23]	26.97	67.51	80.56	27.80	31.49	72.74	85.01	22.27
D^2RL[6]	28.90	70.60	82.40	29.20	-	-	-	-
HI-CMD[15]	34.94	77.58	-	35.94	-	-	-	-
HPILN[24]	41.36	84.78	94.51	42.95	47.56	88.13	95.98	36.08
AlignGAN[5]	42.40	85.00	93.70	40.70	51.50	89.40	95.70	33.90
AGW[7]	47.50	-	-	47.65	-	-	-	-
DDAG[8]	54.75	90.39	95.81	53.02	-	-	-	-
DGTL[25]	57.34	-	-	55.13	-	-	-	-
cm-SSFT[26]	61.60	89.20	93.90	63.20	63.40	91.20	95.70	62.00
HcTri[20]	61.68	93.10	97.17	57.51	-	-	-	-
MSA[28]	63.13	-	-	59.22	-	-	-	-
SFANet[29]	65.74	92.98	97.05	60.83	-	-	-	-
MPANet[9]	70.58	96.21	98.80	68.24	75.58	97.91	99.43	62.91
Ours	72.96	96.67	98.82	69.37	78.50	97.67	99.19	63.99
Indoor Search								
Zero-Pad[13]	20.58	68.38	85.79	26.92	24.43	75.86	91.32	18.64
cmGAN[23]	31.63	77.23	89.18	42.19	37.00	80.94	92.11	32.76
HPILN[24]	45.77	91.82	98.46	56.52	53.50	93.71	98.93	47.48
AlignGAN[5]	45.90	87.60	94.40	54.30	57.10	92.70	97.40	45.30
AGW[7]	54.17	-	-	62.97	-	-	-	-
DDAG[8]	61.02	94.06	98.41	67.98	-	-	-	-
DGTL[25]	63.11	-	-	69.20	-	-	-	-
cm-SSFT[26]	70.50	94.90	97.70	72.60	73.00	96.30	99.10	72.40
HcTri[20]	63.41	91.69	95.28	68.17	-	-	-	-
MSA[28]	67.18	-	-	72.74	-	-	-	-
SFANet[29]	71.60	96.60	99.45	80.05	-	-	-	-
MPANet[9]	76.74	98.21	99.57	80.95	84.22	99.66	99.96	75.11
Ours	77.36	97.76	99.34	80.87	85.73	99.30	99.88	76.07

are plotted as histograms as shown in Fig 6, where Fig 6(a) is using heterocentric sample loss and Fig 6(b) is using heterocentric triplet loss. We calculated their means and variances, where means for (a) are 12.16 with a variance of 2.34 and means for (b) are 12.96 with a variance of 3.17. This suggests that heterocentric sample loss can make the intra-class features more compact.

4.3 Comparison with the state-of-the-art methods.

In this section, we compare our approach with some SOTA methods on SYSU-MM01 and RegDB. The results are shown in Table 3 and Table 4, respectively. The compared methods include Zero-Pad [13], D^2RL [6], cmGAN [23], HI-CMD [15], HPILN [24], AlignGAN [5], DDAG [8], AGW [7], DGTL [25], cm-SSFT [26], MPMN [27], MSA [28], SFANet [29], HAT [30], HcTri [20], MPANet [9]. The results of the comparison method are directly taken from the original article, where "-" means that the corresponding result is not reported in the corresponding article.

Table 4. Comparison with state-of-the-art methods on the REGDB dataset with different query settings. 1^{st} and 2^{nd} best results are marked in red and blue, respectively.

Method	R1	R10	R20	mAP
Visible to Infrared				
Zero-Pad[13]	17.75	34.21	44.35	18.90
D^2RL[6]	43.40	-	-	44.10
AlignGAN[5]	57.90	-	-	53.60
DDAG[8]	69.34	86.19	91.49	63.46
AGW[7]	70.05	-	-	66.37
HAT[30]	71.83	87.16	92.16	67.56
cm-SSFT[26]	72.30	-	-	72.90
SFANet[29]	76.31	91.02	94.27	68.00
MPANet[9]	83.70	-	-	80.90
DGTL[25]	83.92	-	-	73.78
MPMN[27]	86.56	96.68	98.28	82.91
HcTri[20]	91.05	97.16	98.57	83.28
Ours	85.79	99.80	100.00	81.91
Infrared to Visible				
Zero-Pad[13]	16.63	34.68	44.25	17.82
AlignGAN[5]	56.30	-	-	53.40
DDAG[8]	68.06	85.15	90.31	61.80
HAT[30]	70.02	86.45	91.61	66.30
SFANet[29]	70.15	85.24	89.27	63.77
DGTL[25]	81.59	-	-	71.65
MPANet[9]	82.80	-	-	80.70
MPMN[27]	84.62	95.51	97.33	79.49
HcTri[20]	89.30	96.41	98.16	81.46
Ours	86.33	99.41	99.80	82.10

As can be seen from Table 3, on the SYSU-MM01 dataset, our method improves the Rank1 and mAP of both single and multiple searches in all search mode compared to the highest-performing MPANet. In the indoor search mode, the mAP of only single search is slightly lower than MPANet by 0.08, and the remaining Rank1 and mAP are 1.03 higher than MPANet on average.

Table 4 shows the comparison results on the RegDB dataset. It can be seen that the performance of our method reaches a high level compared to current mainstream methods. Among them, the mAP, Rank10, and Rank20 of our method reach the highest level under the search from infrared to visible. HcTri reaches the highest Rank1 and mAP on RegDB. This is because the HcTri is more focused on extracting spatial features. Meanwhile, in the RegDB dataset, there are many image pairs with one-to-one correspondence of spatial locations. Therefore HcTri network is more advantageous. In the SYSU-MM01 dataset, which is more in line with realistic scenarios, the results of our method are much higher than those of HcTri.

5 Conclusion

In this paper, we propose a novel SCFNet for VI-ReID tasks. It mines the feature information of the person in spatial and channel dimensions. To motivate the model to learn color-independent features, we use a random channel enhancement method. Also, a heterocentric sample loss optimization network training process is introduced to make the person feature more compact and distinguishable. Many experiments were conducted on SYSU-MM01 and RegDB to demonstrate the effectiveness of our proposed method. In the future, we will also explore the generalizability of the method and its performance in single-modality person ReID.

Acknowledgements This work was supported by the Project of NSFC (Grant No. U1908214, 61906032), Special Project of Central Government Guiding Local Science and Technology Development (Grant No. 2021JH6/10500140), the Program for Innovative Research Team in University of Liaoning Province (LT2020015), the Support Plan for Key Field Innovation Team of Dalian(2021RT06), the Science and Technology Innovation Fund of Dalian (Grant No. 2020JJ25CY001), the Support Plan for Leading Innovation Team of Dalian University (Grant No. XLJ202010), the Fundamental Research Funds for the Central Universities (Grant No. DUT21TD107), Dalian University Scientific Research Platform Project(No. 202101YB03)

References

1. Sun, Y., Zheng, L., Li, Y., Yang, Y., Tian, Q., Wang, S.: Learning part-based convolutional features for person re-identification. IEEE transactions on pattern analysis and machine intelligence **43**(3), 902–917 (2019). https://doi.org/10.1109/TPAMI.2019.2938523

2. 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). https://doi.org/10.1145/3240508.3240552
3. Xia, B.N., Gong, Y., Zhang, Y., Poellabauer, C.: Second-order non-local attention networks for person re-identification. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3760–3769 (2019). https://doi.org/10.1109/ICCV.2019.00386
4. Zheng, L., Huang, Y., Lu, H., Yang, Y.: Pose-invariant embedding for deep person re-identification. IEEE Transactions on Image Processing **28**(9), 4500–4509 (2019). https://doi.org/10.1109/TIP.2019.2910414
5. 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. pp. 3623–3632 (2019). https://doi.org/10.1109/ICCV.2019.00372
6. Wang, Z., Wang, Z., Zheng, Y., Chuang, Y.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. pp. 618–626 (2019). https://doi.org/10.1109/CVPR.2019.00071
7. Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., Hoi, S.C.: Deep learning for person re-identification: A survey and outlook. IEEE transactions on pattern analysis and machine intelligence **44**(6), 2872–2893 (2021). https://doi.org/10.1109/TPAMI.2021.3054775
8. Ye, M., Shen, J., J Crandall, D., Shao, L., Luo, J.: Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In: European Conference on Computer Vision. pp. 229–247. Springer (2020). https://doi.org/10.1007/978-3-030-58520-4_14
9. Wu, Q., Dai, P., Chen, J., Lin, C.W., 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. pp. 4330–4339 (2021). https://doi.org/10.1109/CVPR46437.2021.00431
10. Ye, M., Lan, X., Wang, Z., Yuen, P.C.: Bi-directional center-constrained top-ranking for visible thermal person re-identification. IEEE Transactions on Information Forensics and Security **15**, 407–419 (2019). https://doi.org/10.1109/TIFS.2019.2921454
11. 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). https://doi.org/10.1016/j.neucom.2019.12.100
12. Li, W., Qi, K., Chen, W., Zhou, Y.: Unified batch all triplet loss for visible-infrared person re-identification. In: 2021 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2021). https://doi.org/10.1109/IJCNN52387.2021.9533325
13. Wu, A., Zheng, W.S., Yu, H.X., Gong, S., Lai, J.: Rgb-infrared cross-modality person re-identification. In: Proceedings of the IEEE international conference on computer vision. pp. 5380–5389 (2017). https://doi.org/10.1109/ICCV.2017.575
14. Wang, G.A., Zhang, T., Yang, Y., Cheng, J., Chang, J., Liang, X., Hou, Z.G.: Cross-modality paired-images generation for rgb-infrared person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 12144–12151 (2020). https://doi.org/10.1609/aaai.v34i07.6894

15. 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). https://doi.org/10.1109/CVPR42600.2020.01027
16. Hao, Y., Li, J., Wang, N., Gao, X.: Modality adversarial neural network for visible-thermal person re-identification. Pattern Recognition **107**, 107533 (2020). https://doi.org/10.1016/j.patcog.2020.107533
17. 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). https://doi.org/10.1016/j.neucom.2020.01.089
18. Huang, N., Liu, J., Zhang, Q., Han, J.: Exploring modality-shared appearance features and modality-invariant relation features for cross-modality person re-identification. arXiv preprint arXiv:2104.11539 (2021). https://doi.org/10.48550/arXiv.2104.11539
19. Zhang, C., Liu, H., Guo, W., Ye, M.: Multi-scale cascading network with compact feature learning for rgb-infrared person re-identification. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 8679–8686. IEEE (2021). https://doi.org/10.1109/ICPR48806.2021.9412576
20. Liu, H., Tan, X., Zhou, X.: Parameter sharing exploration and hetero-center triplet loss for visible-thermal person re-identification. IEEE Transactions on Multimedia **23**, 4414–4425 (2020). https://doi.org/10.1109/TMM.2020.3042080
21. Sun, Y., Zheng, L., Yang, Y., Tian, Q., Wang, S.: Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In: Proceedings of the European conference on computer vision (ECCV). pp. 480–496 (2018). https://doi.org/10.48550/arXiv.1711.09349
22. Nguyen, D.T., Hong, H.G., Kim, K.W., Park, K.R.: Person recognition system based on a combination of body images from visible light and thermal cameras. Sensors **17**(3), 605 (2017). https://doi.org/10.3390/s17030605
23. Dai, P., Ji, R., Wang, H., Wu, Q., Huang, Y.: Cross-modality person re-identification with generative adversarial training. In: IJCAI. vol. 1, p. 6 (2018). https://doi.org/10.24963/ijcai.2018/94
24. Zhao, Y.B., Lin, J.W., Xuan, Q., Xi, X.: Hpiln: a feature learning framework for cross-modality person re-identification. IET Image Processing **13**(14), 2897–2904 (2019). https://doi.org/10.1049/iet-ipr.2019.0699
25. Liu, H., Chai, Y., Tan, X., Li, D., Zhou, X.: Strong but simple baseline with dual-granularity triplet loss for visible-thermal person re-identification. IEEE Signal Processing Letters **28**, 653–657 (2021). https://doi.org/10.1109/LSP.2021.3065903
26. 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. pp. 13379–13389 (2020). https://doi.org/10.1109/CVPR42600.2020.01339
27. Wang, P., Zhao, Z., Su, F., Zhao, Y., Wang, H., Yang, L., Li, Y.: Deep multi-patch matching network for visible thermal person re-identification. IEEE Transactions on Multimedia **23**, 1474–1488 (2020). https://doi.org/10.1109/TMM.2020.2999180
28. Miao, Z., Liu, H., Shi, W., Xu, W., Ye, H.: Modality-aware style adaptation for rgb-infrared person re-identification. In: IJCAI. pp. 916–922 (2021). https://doi.org/10.24963/ijcai.2021/127

29. Liu, H., Ma, S., Xia, D., Li, S.: Sfanet: A spectrum-aware feature augmentation network for visible-infrared person reidentification. IEEE Transactions on Neural Networks and Learning Systems (2021). https://doi.org/10.1109/TNNLS.2021.3105702
30. Ye, M., Shen, J., Shao, L.: Visible-infrared person re-identification via homogeneous augmented tri-modal learning. IEEE Transactions on Information Forensics and Security **16**, 728–739 (2020). https://doi.org/10.1109/TIFS.2020.3001665