Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Multi-granularity Feature Utilization Network for Cross-modality Visible-Infrared Person Re-identification

Guoqing Zhang
Nanjing University of Information Science and Technology
Yinyin Zhang (mkkdzz2428@163.com)
Nanjing University of Information Science and Technology https://orcid.org/0000-0001-8075-3552
Yuhao Chen
Nanjing University of Information Science and Technology
Hongwei Zhang
Nanjing University of Information Science and Technology
Yuhui Zheng
Nanjing University of Information Science and Technology

Research Article

Keywords: Cross-modality , Person re-identification , Multi-granularity , Heterogeneous center loss

Posted Date: May 10th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1458325/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Multi-granularity Feature Utilization Network for Cross-modality Visible-Infrared Person Re-identification

Guoqing Zhang[1*], Yinyin Zhang[1], Yuhao Chen[1], Hongwei Zhang[1] and Yuhui Zheng[1]

[1*]School of Computer Science, Nanjing University of Information Science and Technology Nanjing, 210044, China.

*Corresponding author(s). E-mail(s): xiayang14551@163.com;
Contributing authors: mkkdzz2428@163.com;

Abstract

Cross-modality Visible-Infrared Person Re-identification (VI-ReID) aims to recognize images with the same identity between visible modality and infrared modality, which is a very challenging task. Because it not only includes the trouble of variations between cross-cameras in traditional person re-identification, but also suffers from the huge differences between two modalities. Some existing VI-ReID methods are often limited to single learning of global features or local features, while ignoring the complementarity between fine-grained and coarse-grained information. To solve this problem, our paper designs a multi-granularity feature utilization network (MFUN), which makes up for the lack of modality-shared information by learning fine-grained and coarse-grained features. Firstly, for the sake of learning fine-grained information, a local feature constraint module is introduced, in this module we use hard-mining triplet loss and heterogeneous center loss to constrain local features simultaneously to better promote intra-class closeness and inter-class differences at the coarse and fine granularity levels. Then, our method uses a multi-modality feature aggregation module for global features to fuse the information of two modalities to narrow the modality gap. Through the combination of these two modules, visible and infrared image features can be better fused, thus alleviating the problem of modality discrepancy and supplementing the lack of modality-shared information. Extensive experimental results on RegDB and SYSU-MM01 datasets fully prove that our proposed MFUN has superiority over the state-of-the-art solutions.

Keywords: Cross-modality · Person re-identification · Multi-granularity · Heterogeneous center loss

1 Introduction

Person re-identification (ReID) is among the most popular research directions in the realm of computer vision (Zheng et al. 2016), and its goal is to recognize pedestrians with the same identity in non-overlapping cross-cameras. The main challenge of solving ReID task is to overcome the variations between cameras, such as the viewpoint, pose, illumination, background clutter and occlusion (Leng et al. 2020; Tang et al. 2019; Zhang et al. 2021b). In recent years, a great many models for Re-ID have emerged, most of them mainly deal with the problem as a single-modality retrieval task in visible scene (Ding et al. 2015; Varior et al. 2016; Xiao et al. 2016; Kulis et al. 2011; Liao et al. 2015; Chen et al. 2021; Zhang et al. 2021c; 2021a; Kumar et al. 2020), such as occluded re-ID (Chen et al. 2021), cross-resolution re-ID (Zhang et al.

2021c), video-based re-ID (Zhang et al. 2021a) and unsupervised re-ID (Kumar et al. 2020). However, with the demand of public safety, infrared cameras are widely used in poor-illumination conditions (e.g. in the evening or rayless indoors), and visible-infrared person re-identification (VI-ReID) became a clamant problem, attracting more and more attention.

The purpose of VI-ReID is to search corresponding infrared (visible) images from the gallery set according to the given visible (infrared) image of a specific identity (Wu et al. 2017). This is a very challenging problem, because it will not only face the problems in traditional single-modality ReID, but also be interfered by the cross-modality discrepancies between heterogeneous images. Cross-modality discrepancies mainly come from the inherent different imaging processes of visible and infrared cameras. Figure 1 shows some heterogeneous images taken by different spectral cameras in SYSU-MM01 dataset (Wu et al. 2017). Through observation, it is found that the images taken by infrared cameras lose color information and have exposure problem, which make the traditional ReID method difficult to apply to VI-ReID task.

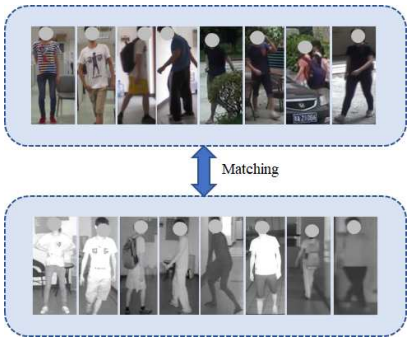

Fig. 1 Some examples in SYSU-MM01 dataset.

To solve VI-ReID task, previous papers have proposed a series of methods, which can be broadly classified into two types: (1) Feature extraction based (Wu et al. 2017; Ye et al. 2020b; Feng et al. 2020; Ye et al. 2018b; Zhu et al. 2020; Zhang et al. 2021), (2) Image generation based (Dai et al. 2018; Ye et al. 2019; Wang et al. 2019a; Choi et al. 2020; Li et al. 2018). For these feature extraction based methods, which usually use a one-stream or two-stream structure to extract modality-specific information and modality-shared information for two modalities. However, they only take the modality independence in global features into consideration, while neglect the modality differences in local fine-grained features, which leads to poor cross-modality generalization capability of VI-ReID models.

From Zhu et al. (2020) and Sun et al. (2021), we found the importance of alleviating cross-modality differences for dealing with VI-ReID task. The difference is that Zhu et al. (2020) measures fine-grained information in local features of pedestrians to reduce cross-modality variation and improve intra-class cross-modality similarity, and Sun et al. (2021) mainly deals with large inter-modality differences through modality relations in each channel of global features. Both works are limited to single learning of local features or global features, without paying attention to the complementarity of information contained in local and global features. Therefore, we integrate the two tasks into a network, which is compatible with the learning of global and local features, so as to better alleviate the differences between modes and supplement the missing modality-shared information.

In view of the above problem, our paper introduces a multi-granularity feature utilization network (MFUN), which makes up for the lack of modality-shared information by learning fine-grained and coarse-grained features. To learn fine-grained information in local features, our method uses a feature extraction module with two independent feature extractors to extract two modalities' local features respectively, and then projects them into the local feature constraint module (Zhu et al. 2020). To constrain local features, cross entropy loss, hard mining triplet loss and heterogeneous center loss (Zhu et al. 2020; Liu et al. 2020; 2021) are used in local feature constraint module. Among them, heterogeneous center loss is center-based loss, and hard mining triplet loss is sample-based. Using these two loss functions concurrently can better promote intra-class compactness and inter-class differences at coarse-grained and fine-grained levels, respectively.

In addition, different from other similar methods (Wei et al. 2020b; Xiang et al. 2019), our paper does not directly process the global features extracted from images, while concatenating the features in local feature constraint module as

global features, and then inputs them into the subsequent modules. Our method also uses a multi-modality feature aggregation module for global features to mix the information of two modalities to narrow the gap and supplement the shared information, so as to relieve the problem of lacking modality-shared information in heterogeneous images.

The main contributions of our paper can be generalized as follows:

1. We design a multi-granularity feature utilization network (MFUN), which pays attention to both local features and global features, and can make up for the shortage of modality-shared information.
2. We combine coarse-grained heterogeneous center loss with fine-grained hard-mining triplet loss for the metric learning of local features to better promote intra-class closeness and inter-class differences.
3. Sufficient experiments on RegDB and SYSU-MM01 datasets prove that our proposed MFUN has superiority over the state-of-the-art solutions.

2 Related works

2.1 Person re-identification

Person Re-identification (ReID), as a key component in the domain of video surveillance, has gradually becomes a research hotspot in recent years. Its purpose is to accurately identify a certain target pedestrian in the sight of surveillance cameras and pedestrians in the sight of other surveillance cameras. However, in actual monitoring, environmental changes are often uncontrollable, which lead to a series of complex problems, such as viewpoint, posture change, background clutter and occlusion. Considering the interference of these problems, some existing methods pay attention to exacting robust feature representations or learning discriminative distance metrics to improve the performance by using predefined positions (Zhang et al. 2017) and semantic regions (Wei et al. 2018; Zhang et al. 2021d), including local stripes, pedestrian parts and attention mechanisms. These existing ReID solutions have surpassed the performance of human level, but majority of them are proposed for single-modality ReID. However, in the night or dark environment, visible cameras will be replaced by infrared cameras, and the images taken by infrared cameras is grayish white, which is quite different from visible images, traditional ReID methods are not applicable. Consequently, it is of great importance to handle the cross-modality ReID task.

2.2 Visible-infrared person re-identification

Visible-Infrared Person Re-identification (VI-ReID) attempts to identify person images with same identity between two modalities, it has received increasing attention recently. Wu et al. (2017) defined cross-modality ReID task for the first time, and contribute a standard dataset named SYSU-MM01 to support its research. To solve this task, they analyzed the differences among three commonly used cross-domain network architectures and further exploit a one-stream network architecture, namely zero-padding network. Afterwards, Ye et al. (2018a) indicated that the key to deal with cross-modality ReID is to solve cross-modality and intra-modality variations simultaneously. On account of this viewpoint, they designed a metric learning method called HCML. HCML is realized by jointly optimizing modality-specific and modality-shared features. In this method, two modalities' features are projected into a subspace to minimize the cross-modality and intra-modality changes. Additionally, Ye et al. (2020b) presented a dual-stream structure called TONE. TONE is trained by joint supervision of contrast loss and cross entropy loss in the training stage.

In addition, some works have applied GAN to VI-ReID task. Dai et al. (2018) introduced a method called cmGAN. The method uses the idea of confrontation training of generators and discriminators in GAN. This network uses the generator to learn the features of different modalities. The purpose of discriminator is to differentiate whether the input features come from visible modalities or infrared modality. Li et al. (2020) introduced a noval X-modality to reduce modality differences, and proposes a XIV-ReID method, which redefines infrared-visible cross-modality problem as an X-infrared-visible tri-modality task. The generator extracts information

from visible and infrared images to generate X-modality images. Then, cross-modality information of three modalities is learned in a common space.

However, the VI-ReID task is still very challenging because of the great changes in vision. Although previous studies have greatly improved the accuracy of VI-ReID, they are still far from satisfactory. Some existing methods are limited to single learning of local features or global features to obtain distinctive features, while ignoring the complementarity between fine-grained and coarse-grained information. Therefore, our paper proposes a multi-granularity feature utilization network (MFUN), in which modality-shared features in fine-grained and coarse-grained information are better paid attention to by combining local embedding module and multi-modality feature aggregation module.

3 The Proposed Method

3.1 Network structure

The framework of MFUN is shown in Figure 2, which is made up of three parts: feature extraction module, local feature constraint module and multi-modality feature aggregation module.

We express the input visible set and infrared set as $X^V = \{x_i^v | x_i^v \in \mathbb{R}^{C\times H\times W}\}$ and $X^T = \{x_i^t | x_i^t \in \mathbb{R}^{C\times H\times W}\}$.The C,H,W denote the channel, height and width of the images. There are K images with the same number of x_i^v and x_i^t in each training batch, where $i = 1, 2, ..., K/2$. During training, x_i^v and x_i^t in each input batch are fed into the corresponding branches. Firstly, the local features of the two modalities are extracted and defined as follows:

$$V_P = \phi^v\{x_i^v\} \qquad T_P = \phi^t\{x_i^t\} \tag{1}$$

in which $P = 1, 2, ..., p$, ϕ^v and ϕ^t represent feature extraction module of visible and infrared modalities respectively.

After that, L2 regularization and FC layer are used to project the local features V_P and T_P as $F_1, F_2, ..., F_p$, and they are fed into a common subspace, and constrained by three loss functions for joint training of the local feature constraint module. Then, p local features $F_1, F_2, ..., F_p$ are concatenated as global feature F, which is input into the multi-modality feature aggregation module to calculate its relation matrix to update the global feature. Finally, we use cross-entropy loss and hard-mining triplet loss to constrain the updated global features.

3.2 Feature extraction module

The functions of two feature extractors in feature extraction module are to extract modality-specific features (such as colors) and modality-shared features(such as contours and textures) of two modality images respectively. Modality-specific features exist only in a specific modality, or changes with the variations of modality, which will enlarge the distance between images with the same identity in two modalities. Modality-shared features are invariant information between two modalities, which are robust to modality variations and are beneficial to VI-ReID task.

In our paper, the feature extraction module uses two branches, which are independent of each other and have the same structure. First of all, two modalities' images should be input into the corresponding feature extractors. Figure 3 shows the details of a feature extractor in one branch. As shown in the figure, the backbone network used by feature extractor is ResNet-50. The difference is that in order to enrich the granularity of features and keep more fine-grained features, we removed the last down-sampling operation in ResNet-50 (Zhu et al. 2020). According to Part-Based Convolutional Baseline (PCB) (Sun et al. 2018), we divide the extracted modality-shared information into p horizontal stripes with global average pooling (GAP), and each stripe is averaged into a local feature vector. Then, a Leaky ReLU activation layer and a batch normalization (BN) layer are used to reduce the dimension of each local feature vector. Because the feature extraction module uses two independent branches, the local features of visible and infrared modalities are independent of each other. Therefore, L2 regularization and FC layer are used to project them into a common feature subspace for subsequent joint training of local feature constraint module.

3.3 Local feature constraint module

The purpose of local feature constraint module is to enlarge the differences between classes and improve the inter-modality similarity within a

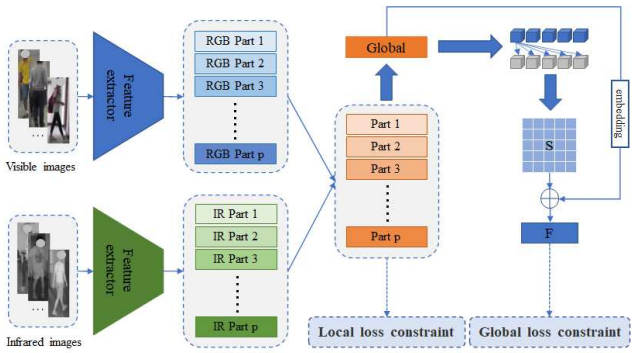

Fig. 2 Overall network architecture.

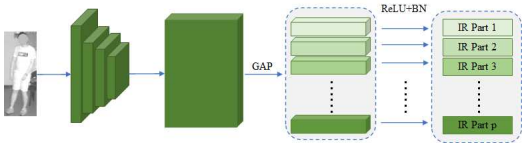

Fig. 3 Details of feature extractor.

class. In our method, heterogeneous center loss, hard-mining triplet loss and cross-entropy loss are used to supervise the training of local features (Liu et al. 2021).

Heterogeneous center loss is center-based triplet loss, and hard-mining triplet loss is based on samples. They can reduce the distance between two modalities features within a class and widen the gap between classes at coarse-grained and fine-grained levels, respectively.

Heterogeneous center loss is used to improve inter-modality similarity within a class. Intuitively, it is by punishing the center distance between two modality distributions. The center of each class is directly calculated based on the extracted features. In each mini-batch, the calculation formula of feature center of each identity of each modality is as follows (Zhu et al. 2020):

$$c_v^i = \frac{1}{K}\sum_{m=1}^{K} x_{i,m}^v \tag{2}$$

$$c_t^i = \frac{1}{K}\sum_{m=1}^{K} x_{i,m}^t \tag{3}$$

Where $x_{i,m}$ represents the m-th image feature of the i-th pedestrian, v and t correspond to visible and infrared modalities.

Heterogeneous center loss used in our paper is to directly calculate the distance between two centers, and L_{hc} can be computed by:

$$L_{hc} = dist(c_v^i - c_t^i) \tag{4}$$

Where $dist(.)$ represents the cosine distance.

The function of heterogeneous center loss is only to close the center distance of each class, so as to promote the inter-modality similarity within a class, and there is no learned distinguishing feature representation to enlarge the differences between classes. Therefore, our method also uses hard-mining triplet loss to constrain local features. For each feature F_p^a in the mini-batch, the hardest positive F_p^p and hardest negative F_p^n can be mined to form a triplet, and the fine-grained triplet function is as follows:

$$L_{tri\_p} = \sum_{i=1}^{P}\sum_{a=1}^{2K}[m + max\|F_P^{a,i} - F_P^{p,i}\|_2 - min\|F_P^{a,i} - F_P^{n,j}\|_2]_+ \tag{5}$$

m represents margin, and $F_P^{a,i}$ represents the P-th local feature of the a-th image of the i-th person, $\|.\|_2$ represents the Euclidean distance between two feature vectors, $[.]_+ = max(.,0)$.

In addition to the hard-mining triplet loss, we adopt the cross-entropy loss to enhance the learning of distinguishing features together with it as well. The formula of the cross-entropy loss can be represented as:

$$L_{id\_p} = -\sum_{i=1}^{K}\log\frac{e^{W_{y_i}^T x_i^a + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T x_i^a + b_m}}, a \in [1,p] \tag{6}$$

K represents batch size, x_i^a is the feature vector of i-th image from a-th part's feature map, y_i is the

identity of the i-th person, W_j is j-th identity's classifier, and b is bias term.

Therefore, the total local loss constraint is

$$L_{local} = L_{id\_p} + \lambda_1 L_{hc} + \lambda_2 L_{tri\_p} \tag{7}$$

λ_1 and λ_2 are weight parameters to determine the two kinds of triplet loss used in the local feature constraint module.

3.4 Multi-modality feature aggregation module

While using local information, our paper also pays attention to global information. Different from other methods, our method fuses p local features in the common subspace, and regards the fused features as global features (Liu et al. 2020). For global features, in order to supplement the missing information in a single modality, we input them into a multi-modality feature aggregation module (Sun et al. 2021). This module is used to fuse cross-modality features. It uses cross-modality relations to update the original global features to narrow the huge modality difference.

In the multi-modality feature aggregation module, firstly, the relation matrix S of global features is calculated. Global features need to be divided into visible features F^R and infrared features F^I, which are reduced the dimension by a 1×1 spatial convolution layer, followed by BN layer and ReLU activation layer , so that they can pass more effective modality information. After getting the embedded features F^R and F^I, the feature map at each channel is regarded as a feature vector, and the paired Euclidean distance between each feature vector in F^R and all feature vectors in F^I is calculated, then the relation matrix S can be obtained.

$$S = \begin{pmatrix} d(F_1^R, F_1^I) & \cdots & d(F_1^R, F_c^I) \\ \vdots & \ddots & \vdots \\ d(F_c^R, F_1^I) & \cdots & d(F_c^R, F_c^I) \end{pmatrix} \tag{8}$$

$$d(F_i^R, F_j^I) = 1 - 0.5 \| \frac{F_i^R}{\|F_i^R\|} - \frac{F_j^I}{\|F_j^I\|} \| \tag{9}$$

To avoid losing the original features, the relation matrix S is combined with the original feature to update the global feature F.

$$F = sigmoid(W[\phi(F), \varphi(S)]) \tag{10}$$

Where ϕ and φ represent embedded operations for original and relational features, F and S represent original feature and relational feature respectively, and W represents learnable parameter.

For the updated global features, following the common feature learning strategy, our paper adopts cross-entropy loss and hard-mining triplet loss to improve distinctiveness of global features. The formulas are as follows.

$$L_{id\_g} = -\sum_{i=1}^{K} \log \frac{e^{W_{y_i}^T x^i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T x^i + b_m}} \tag{11}$$

$$L_{tri\_g} = \sum_{i=1}^{P} \sum_{a=1}^{2K} [m + max\|F^{a,i} - F^{p,i}\|_2 - min\|F^{a,i} - F^{n,j}\|_2]_+ \tag{12}$$

In cross-entropy loss L_{id_g}, x_i denotes the features of the i-th person collected from y_i class, y_i is the identity of the i-th person. In the hard-mining triplet loss L_{tri_g}, $F^{a,i}$ represents the global feature of the a-th image of the i-th person.

Therefore, the total global loss constraint is

$$L_{global} = L_{id\_g} + L_{tri\_g} \tag{13}$$

Finally, the final optimization goal of our network can be written as follows:

$$L_{all} = L_{local} + L_{global} \tag{14}$$

4 Experiment

4.1 Experimental settings

Dataset. Our experiments are carried out on two datasets RegDB (Dat et al. 2017) and SYSU-MM01 (Wu et al 2017) to evaluate the effectiveness of our proposed MFUN, these two datasets are common standard benchmarks used by previous VI-ReID solutions.

RegDB dataset is taken by a dual-camera system. It includes 8,024 images of 412 identities, in which there are 10 visible and infrared images of each person, including 254 females and 158 males.

Among the 412 identities, 156 were shot from the front and 256 were shot from the back. The dataset is randomly divided into two parts, each for training and testing. That is to say, training set and testing set have 2,060 visible images and 2,060 infrared images respectively. We set visible images as query set and infrared images as gallery set in the testing stage.

SYSU-MM01 dataset is the largest dataset in visible-infrared cross-modality Re-ID. The dataset is composed of 491 identity images collected by 6 cameras, of which 4 are visible cameras and 2 are infrared cameras. Each pedestrian is observed by at least one visible camera and one infrared camera. The captured images include 287,628 visible images and 15,792 infrared images. Our training set consists of 22,258 visible images and 11,909 infrared images of 395 pedestrians, and testing set consists of 96 identities's visible and infrared images. We use two evaluation modes in SYSU-MM01: indoor-search and all-search. Both of their query set is the same, but for all-search mode, gallery images are the images observed by all the visible cameras, while for indoor-search mode, gallery images are only the images from indoor visible cameras.

Evaluation Metric. The evaluations of our method are carried out by using Cumulative Matching Characteristic curve (CMC) and mean Average Precision (mAP). In our paper, CMC is written as R-k, which measures the rate of a correct image occurs in the k-nearest matching results.

Experimental Settings. Our proposed method is implemented with PyTorch deep learning framework. All the input images are resized to 288×144, besides, we put random horizontal flip and random cropping to use for data augmentation during training phase. Each mini-batch consists of 64 pedestrian images of 4 identities, that is to say, 8 sets of visible and infrared images need to be stochastically selected from each identity. SGD optimizer is adopted in our work, and the momentum parameter is set to 0.9. Our model is trained for 60 epochs in all, and the learning rate of the first 30 epochs is set to 1×10^{-2}, then we decay the learning rate to 1×10^{-4} at the last 30 epochs. The backbone network ResNet-50 used by feature extractor is pretrained on ImageNet. The features extracted by the feature extractor are averagely divided into $p = 6$ horizontal stripes. For the margin parameter m in hard-mining triplet loss, we set it to 0.3. In the total loss function, the weight λ_1 of heterogeneous center loss and the weight λ_2 of hard-mining triplet loss of local features are set to 0.6 and 0.8, respectively.

4.2 Comparison with the state-of-art methods

Next, We will present comparison with several state-of-the-art VI-ReID solutions on SYSU-MM01 and RegDB datasets. The competing methods contain the feature extraction based methods (MAC (Ye et al. 2019), HPILN (Zhao et al. 2019), DEF (Hao et al. 2019), MSPAC-MeCen (Zhang et al. 2020), MACE (Ye et al. 2020a), CMM (Ling et al. 2020), CPN (Zhang et al. 2021e), DDAG (Ye et al. 2020c), HAT (Ye et al. 2021b), FbA (Park et al. 2021), AGW (Ye et al. 2021a), CoAL (Wei et al. 2020a), FBP (Wei et al. 2021), NFS (Chen et al. 2021)) and the image generation based methods (D2RL (Wang et al. 2019b), AlignGAN (Wang et al. 2019a), Hi-CMD (Choi et al. 2020), JSIA (Wang et al. 2020), Xmodal (Li et al. 2020), ADC-Net (Hu et al. 2021)). And the comparisons are provided in Table 1 and Table 2.

SYSU-MM01. By observing the comparison on SYSU-MM01 listed in Table 1, we can get the following conclusions. Our MFUN reaches comparable performance compared to the results obtained by DDAG (Ye et al. 2020c), NFS (Chen et al. 2021), CoAL (Wei et al. 2020a) and CPN (Zhang et al. 2021e), and outperformances the other remaining comparison methods. However, our method performs better than DDAG (Ye et al. 2020c), NFS (Chen et al. 2021), CoAL (Wei et al. 2020a) and CPN (Zhang et al. 2021e) in the more challenging mode all-search.. From Table 1, we can intuitively see that under the all-search mode, compared with DDAG (Ye et al. 2020c), NFS (Chen et al. 2021) and CoAL (Wei et al. 2020a), the method proposed in our paper is superior to rank-1 and mAP in single-shot. Under two evaluation modes, compared with CPN (Zhang et al. 2021e), the three mAP scores of our method is higher than its. The results in Table 1 show that, under the all-search mode, the mAP score of our method achieves 58.14% in single-shot and 51.09% in multi-shot, which exceeds all the methods listed, and is at least increased by 1.23% and

Table 1 Comparison with the state-of-the-art on SYSU-MM01 dataset. The 1st, 2nd and 3rd best results are emphasized with red, green and blue color, respectively.

	All-search								Indoor-search							
	Single-shot				Multi-shot				Single-shot				Multi-shot			
Method	R-1	R-10	R-20	mAP	R-1	R-10	R-20	mAP	R-1	R-10	R-20	mAP	R-1	R-10	R-20	mAP
D^{2}RL (Wang et al. 2019b)	28.90	70.6	82.4	29.20	-	-	-	-	-	-	-	-	-	-	-	-
AlignGAN (Wang et al. 2019a)	42.40	85.00	93.70	40.70	51.50	89.40	95.70	33.90	45.90	87.60	94.40	54.30	57.10	92.70	97.40	45.30
Hi-CMD (Choi et al. 2020)	34.94	77.58	-	35.94	-	-	-	-	-	-	-	-	-	-	-	-
JSIA (Wang et al. 2020)	38.10	80.70	89.90	36.90	45.10	85.70	93.80	29.50	43.80	86.20	94.20	52.90	52.70	91.10	96.40	42.70
Xmodal (Li et al. 2020)	49.92	89.79	95.96	50.73	-	-	-	-	-	-	-	-	-	-	-	-
ADCNet (Hu et al. 2021)	55.90	90.10	96.20	59.60	-	-	-	-	58.80	92.80	97.50	65.60	-	-	-	-
MAC (Ye et al. 2019)	33.26	79.04	90.09	36.22	-	-	-	-	-	-	-	-	-	-	-	-
HPILN (Zhao et al. 2019)	41.36	84.78	94.51	42.95	47.56	88.13	95.98	36.08	45.77	91.82	98.46	56.52	53.05	93.71	98.93	47.48
MSPAC-MeCen (Zhang et al. 2020)	46.62	87.59	95.77	47.26	47.57	87.64	96.11	38.53	51.63	93.48	98.82	61.54	52.81	94.16	99.37	47.09
AGW (Ye et al. 2021a)	47.50	84.39	92.14	47.65	-	-	-	-	54.17	91.14	95.98	62.97	-	-	-	-
DEF (Hao et al. 2019)	48.71	88.86	95.27	48.59	54.63	91.62	96.83	42.14	52.25	89.86	95.85	59.68	59.62	94.45	98.07	50.60
MACE (Ye et al. 2020a)	51.64	87.25	94.44	50.11	-	-	-	-	57.35	93.02	97.47	64.79	-	-	-	-
CMM (Ling et al. 2020)	51.80	92.72	97.71	51.21	56.27	94.08	98.12	43.39	54.98	94.38	99.41	63.70	60.42	96.88	99.50	53.52
FBP (Wei et al. 2021)	54.14	86.04	93.03	50.20	-	-	-	-	-	-	-	-	-	-	-	-
DDAG(Ye et al. 2020c)	54.75	90.39	95.81	53.02	-	-	-	-	61.02	94.06	98.41	67.98	-	-	-	-
HAT(Ye et al. 2021b)	55.29	92.14	97.36	53.89	-	-	-	-	62.10	95.75	99.20	69.37	-	-	-	-
LbA(Park et al. 2021)	55.41	-	-	54.14	-	-	-	-	58.64	-	-	66.33	-	-	-	-
NFS (Chen et al. 2021)	56.91	91.34	96.52	55.45	63.51	94.42	97.81	48.56	62.79	96.53	99.07	69.79	70.03	97.70	99.51	61.45
CoAL (Wei et al. 2020a)	57.22	92.29	97.57	57.20	-	-	-	-	63.86	95.41	98.79	70.84	-	-	-	-
CPN (Zhang et al. 2021e)	57.33	92.62	97.14	56.91	63.08	93.89	97.37	50.65	59.30	94.46	98.38	66.70	66.26	97.44	99.79	58.51
Our work	57.25	93.45	98.13	58.14	61.81	94.67	98.32	51.09	60.66	96.10	98.92	69.11	70.64	97.63	99.37	62.26

2.53% respectively. In indoor-search mode, our method presents an enhancement in multi-shot under indoor-search mode, with an 0.61% rise in rank-1 accuracy and an 0.81% rise in mAP score at least.

RegDB. The results of the RegDB dataset listed in Table 2 show that, our work reaches much higher experimental accuracy on RegDB dataset, the comparison between the results clearly shows great advantages of our method. The rank-1 accuracy and mAP score of our proposed method are 87.23% and 81.70%, which are higher than other methods. Compared with the best performing method NFS (Chen et al. 2021), our method improves 6.69% rank-1 accuracy and 9.6% mAP score.

The results shown in Table 1 and 2 strongly proves our method's effectiveness and generalization.

Table 2 Comparison with the state-of-art on RegDB dataset.

Method	R-1	R-10	R-20	mAP
D^{2}RL(Wang et al. 2019b)	43.40	66.10	76.30	44.1
JSIA (Wang et al. 2020)	48.50	-	-	49.30
AlignGAN (Wang et al. 2019a)	57.9	-	-	53.6
Xmodal (Li et al. 2020)	62.21	83.13	91.72	60.18
Hi-CMD (Choi et al. 2020)	70.93	86.39	-	66.0
ADCNet (Hu et al. 2021)	72.90	-	-	66.50
MAC (Ye et al. 2019)	36.43	62.36	71.63	37.0
HPILN (Zhao et al. 2019)	-	-	-	-
DEF (Hao et al. 2019)	48.71	88.86	95.27	48.59
MSPAC-MeCen(2021) (Zhang et al. 2020)	49.61	72.28	80.63	53.64
MACE (Ye et al. 2020a)	-	-	-	-
CMM (Ling et al. 2020)	59.81	80.39	88.69	60.86
CPN (Zhang et al. 2021e)	68.59	84.81	98.33	69.20
DDAG (Ye et al. 2020c)	69.34	86.19	91.49	63.46
AGW (Ye et al. 2021a)	70.05	-	-	66.37
HAT (Ye et al. 2021b)	71.83	87.16	92.16	67.56
FBP (Wei et al. 2021)	73.98	89.71	93.69	68.24
CoAL (Wei et al. 2020a)	74.12	90.23	94.53	69.87
LbA (Park et al. 2021)	74.17	-	-	67.64
NFS (Chen et al. 2021)	80.54	91.96	95.07	72.10
Our work	**87.23**	**96.75**	**98.50**	**81.70**

4.3 Ablation study

To verify the validity of our proposed MFUN and show the advantage of using heterogeneous center loss and hard-mining triplet loss concurrently, we conduct four groups of ablation experiments on the SYSU-MM01 dataset (using the all-search mode in single-shot). The values of ablation study are listed in Table 3, it can clearly see that all the modules used in MFUN can improve the performance. Specifically, the 'P' represents only use the

local features, 'S' represents the multi-modality feature aggregation module only use the relation matrix, 'S+T' represents multi-modality feature aggregation module combines relation matrix with original feature to update global feature, 'HC loss' represents heterogeneous center loss and 'Tri loss' represents hard-mining triplet loss.

Effectiveness of multi-modality feature aggregation module. For the multi-modality aggregation module, we conducted two groups of ablation experiments, version 1 and 2. In version 1, we learn local features independently. As can be seen from Table 3, rank-1 and mAP only reached 55.48% and 55.53%. Compared with the results of version 5, we can find that it is necessary to supervise local and global features simultaneously. In version 2, the global feature only uses the relation matrix, which is not integrated with the original feature. By comparing version 2 and version 5, we can see that although R-1, R-10 and R-20 in version 5 are at most 0.15% lower than those in version 2, the mAP score is improved by 0.69% after fusing the relation matrix with the original feature, which is enough to show the effectiveness of combining the relation matrix with the original feature to update the global feature.

Effectiveness of using heterogeneous center loss and hard-mining triplet loss simultaneously. By comparing the results of version 3, 4 and 5 in Table 3, we can see that rank-1 and mAP only reach 46.11% and 44.59% when using ID loss and heterogeneous center loss for local features, and when using ID loss and hard-mining triplet loss for local features, Rank-1 and mAP only reached 47.46% and 47.99%, respectively. But after using ID loss, heterogeneous center loss and hard-mining triple loss simultaneously, rank-1 and mAP reached 57.25% and 58.14%. It can be seen from the comparison that when three constraints exist simultaneously, the model's performance increased by 9.79% on Rank-1 and 10.15% on mAP at least. The comparison clearly explains that the simultaneous use of ID loss, heterogeneous center loss and triplet loss can better mine the discriminative person representation, and shorten the distance between two modalities features within a class and widen the gap between classes.

4.4 Discussion on the weighting parameters of loss function

In this subsection, we conduct several experiments on λ_1 and λ_2 which are two weighting parameters in Eq.12, that are used to determine the weight of heterogeneous center loss and hard-mining triplet loss. The experiments are carried out on SYSU-MM01 with all-search single-shot mode, with different values of λ_1 and λ_2, the change trend of rank-1 and mAP are drawn in Figure 4 (a) and (b), respectively. We vary λ_1 from 0.1 to 0.9 and empirically set λ_2= 0.8. As shown in Figure 4(a), in the process of λ_1 increasing from 0.1 to 0.6, rank-1 and mAP are also in an upward trend, when λ_1 is greater than 0.6, the results begin to show a downward trend. Then λ_2 will take a value from 0.1 to 0.9 and we empirically set $\lambda_2 = 0.6$. Figure 4(b) shows that the result rises when λ_2 is between 0.1 and 0.5, then from 0.5 to 0.8, the performance is floating. When λ_2 is greater than 0.8, the result gradually decreases. Therefore, we set $\lambda_1 = 0.6$ and $\lambda_2 = 0.8$ in the other experiments.

4.5 Discussion on the number p of local feature partition

In this subsection, we discuss the number p of sharThe value of p determines the size of local feature partition. For p, we took values of 4, 6 and 8 for three groups of experiments on SYSU-MM01, and the influence on the experimental results can be seen from Figure 5. When $p = 4$, the granularity of local features is not detailed enough, which makes it difficult for the network to pay attention to more details, the result only obtains the rank-1 accuracy = 53.24% and mAP score = 53.61%. When $p = 6$, the performance of our method achieves the best, specifically, it achieves the rank-1 accuracy = 57.25% and mAP score = 58.14%. When $p = 8$, the experimental result drops sharply again. Because the granularity is too small, which makes it hard for feature extractors to capture effective local information. By comparison, it can be clearly seen that when $p = 6$, the experimental performance meets the best.

Table 3 Ablation study of different parts of our method on SYSU-MM01 dataset.

Version	P	S	S+T	HC loss	Tri loss	R-1	R-10	R-20	mAP
version 1	✓			✓	✓	55.48	92.24	97.32	55.53
version 2	✓	✓		✓	✓	57.35	93.60	98.28	57.45
version 3	✓	✓	✓	✓		46.11	88.00	94.66	44.59
version 4	✓	✓	✓		✓	47.46	87.50	95.03	47.99
version 5	✓	✓	✓	✓	✓	57.25	93.45	98.13	58.14

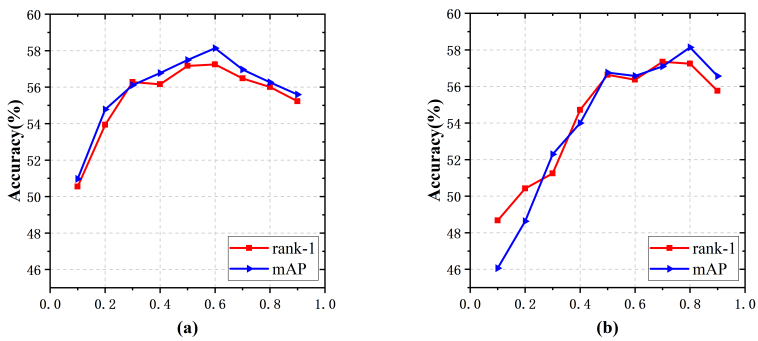

Fig. 4 Evaluation of the weighting parameters λ_1 and λ_2 on SYSU-MM01 dataset. (a) Impact of λ_1. (b) Impact of λ_2.

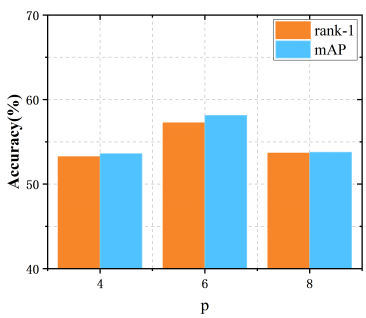

Fig. 5 Impact of the number p of local feature partition

4.6 Visualization of feature distribution

So as to more clearly understand the validity of our proposed method, we visualize the feature distributions on SYSU-MM01 by t-SNE (Van et al. 2008). As shown in Figure 6, (a)-(d) show the visualization of the Initial feature distribution, local branch with heterogeneous center loss and cross-entropy loss, local branch with triplet loss and cross-entropy loss and our method. Various colors represent different identities.

From Figure 6(a), we can observe that the initial feature distributions are disordered, and the features with the same identity are difficult to be well aggregated. In (b) and (c), separately using heterogeneous center loss and triplet loss for local features has played a good role in feature aggregation of the same identity, but compared with (d), it can be clearly observed from (b) and (c) that the feature distributions of the same identity in the red circle is loose and the differences between classes are not well expanded. And from (b), we also observe that many scattered features are not gathered together, which shows that it is not enough to narrow the intra-class gap only by coarse-grained heterogeneous center loss. Compare the first three images, (d) shows that using heterogeneous center loss and triplet loss to constrain local features can effectively promote intra-class closeness and uniformly lager inter-class differences, thereby improving the distinctiveness of features.

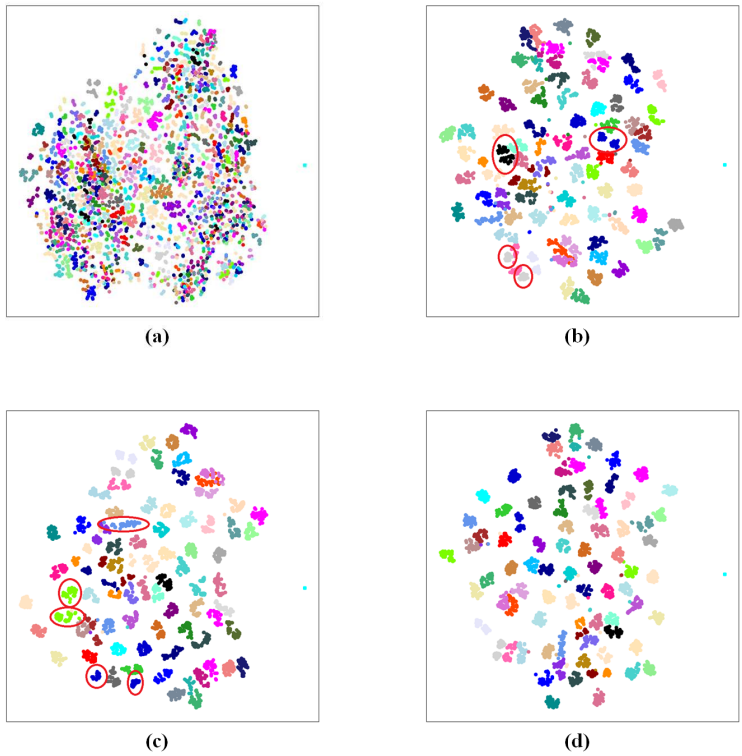

Fig. 6 Feature distributions visualized with t-SNE method. (a) Initial feature distribution. (b) Feature distribution of local branch with heterogeneous center loss and cross-entropy loss. (c) Feature distribution of local branch with triplet loss and cross-entropy loss. (d) Feature distribution of our method.

5 Conclusion

In this paper, we present a novel network for visible-infrared person re-identification, multi-granularity feature utilization network (MFUN), which can take the learning of local and global features into consideration so as to alleviate the problem of missing shared information between cross-modalities. Our method makes up of three parts: feature extraction module, local feature constraint module and multi-modality feature aggregation module, in which local feature constraint module and multi-modality feature aggregation module are used to focus on local and global features, respectively. On one hand, for the sake of better promoting intra-class closeness and inter-class differences at coarse granularity and fine granularity levels, we adopt heterogeneous center loss and hard-mining triplet loss in local feature constraint module to jointly learn local features. On the other hand, in order to improve cross-modality correlation, multi-modality aggregation module is used to weaken modality discrepancies by using global features. Extensive experiments conducted on the frequently-used dataset SYSU-MM01 and RegDB have demonstrated the superiority of our MFUN compared with the state-of-the-art VI-ReID methods.

Author Contributions All authors contributed to the conceptualization and methodology. The writing- original draft, writing-review editing and visualization were performed by Guoqing Zhang and Yinyin Zhang. Investigation and data curation were performed by Yuhao Chen and Hongwei Zhang. Supervision was performed by Yuhui Zheng, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding This work is supported by the National Natural Science Foundation of China (No.

61806099, No. 62172231 and No. U20B2065) and the Natural Science Foundation of Jiangsu Province of China (No. BK20180790 and No. BK20211539).

Declarations

Compliance with Ethical Standards This study does not contain any studies with human participants or animals performed by any of the authors.

Ethical approval Any of the authors' investigations with human participants or animals are not included in this article.

Conflict of interest All the authors declare no conflict of interest.

Informed consent Informed consent was obtained from all individual participants included in the study.

References

Chen Y, Wan L, Li Z, et al (2021) Neural feature search for rgb-infrared person re-identification. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 587–597, https://doi.org/10.1109/CVPR46437.2021.00065

Choi S, Lee S, Kim Y, et al (2020) Hi-cmd: Hierarchical cross-modality disentanglement for visible-infrared person re-identification. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 10,254–10,263, https://doi.org/10.1109/CVPR42600.2020.01027

Dai P, Ji R, Wang H, et al (2018) Cross-modality person re-identification with generative adversarial training. In: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), pp 677–683, https://doi.org/10.24963/ijcai.2018/94

Dat N, Hong H, Ki K, et al (2017) Person recognition system based on a combination of body images from visible light and thermal cameras. Sensors 17(3):605. https://doi.org/10.3390/s17030605

Ding S, Lin L, Wang G, et al (2015) Deep feature learning with relative distance comparison for person re-identification. Pattern Recognition 48(10):2993–3003. https://doi.org/10.1016/j.patcog.2015.04.005

Feng Z, Lai J, Xie X (2020) Learning modality-specific representations for visible-infrared person re-identification. IEEE Transactions on Image Processing 29:579–590. https://doi.org/10.1109/TIP.2019.2928126

Hao Y, Wang N, Gao X, et al (2019) Dual-alignment feature embedding for cross-modality person re-identification. In: The ACM International Conference on Multimedia (ACMMM), pp 57–65, https://doi.org/10.1145/3343031.3351006

Hu B, Liu J, Zha Zj (2021) Adversarial disentanglement and correlation network for rgb-infrared person re-identification. In: IEEE International Conference on Multimedia and Expo (ICME), pp 1–6, https://doi.org/10.1109/ICME51207.2021.9428376

Kulis B, Saenko K, Darrell T (2011) What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 1785–1792, https://doi.org/10.1109/CVPR.2011.5995702

Kumar D, Siva P, Marchwica P, et al (2020) Unsupervised domain adaptation in person re-id via k-reciprocal clustering and large-scale heterogeneous environment synthesis. In: The IEEE Winter Conference on Applications of Computer Vision (WACV), pp 2634–2643, https://doi.org/10.1109/WACV45572.2020.9093606

Leng Q, Ye M, Tian Q (2020) A survey of open-world person re-identification. IEEE Transactions on Circuits and Systems for Video Technology 30(4):1092–1108. https://doi.org/10.1109/TCSVT.2019.2898940

Li D, Wei X, Hong X, et al (2020) Infrared-visible cross-modal person re-identification with an x

modality. In: The AAAI Conference on Artificial Intelligence (AAAI), pp 4610–4617, https://doi.org/10.1609/aaai.v34i04.5891

Liao S, Li SZ (2015) Efficient psd constrained asymmetric metric learning for person re-identification. In: The IEEE International Conference on Computer Vision (ICCV), pp 3685–3693, https://doi.org/10.1109/ICCV.2015.420

Ling Y, Zhong Z, Luo Z, et al (2020) Class-aware modality mix and center-guided metric learning for visible-thermal person re-identification. In: The ACM International Conference on Multimedia (ACMMM), pp 889–897, https://doi.org/10.1145/3394171.3413821

Liu H, Tan X, Zhou X (2020) Parameter sharing exploration and hetero-center triplet loss for visible-thermal person re-identification. IEEE Transactions on Multimedia pp 1–1. https://doi.org/10.1109/TMM.2020.3042080

Liu H, Chai Y, Tan X, et al (2021) Strong but simple baseline with dual-granularity triplet loss for visible-thermal person re-identification. IEEE Signal Processing Letters 28:653–657. https://doi.org/10.1109/LSP.2021.3065903

Van der Maaten L, Hinton G (2008) Visualizing data using t-sne. Journal of Machine Learning Research (JMLR) 9(11)

Park H, Lee S, Lee J, et al (2021) Learning by aligning: Visible-infrared person re-identification using cross-modal correspondences. In: The IEEE International Conference on Computer Vision (ICCV), pp 12,026–12,035, https://doi.org/10.1109/ICCV48922.2021.01183

Sun J, Zhang T (2021) Rgb-infrared person re-identification via multi-modality relation aggregation and graph convolution network. In: IEEE International Conference on Image Processing (ICIP), pp 1174–1178, https://doi.org/10.1109/ICIP42928.2021.9506288

Sun Y, Zheng L, Yang Y, et al (2018) Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In: The European Conference on Computer Vision (ECCV), pp 501–518, https://doi.org/10.1007/978-3-030-01225-0_30

Tang Y, Yang X, Wang N, et al (2019) Person re-identification with gradual background suppression. In: IEEE International Conference on Multimedia and Expo (ICME), pp 706–711, https://doi.org/10.1109/ICME.2019.00127

Varior RR, Shuai B, Lu J, et al (2016) A siamese long short-term memory architecture for human re-identification. In: The European Conference on Computer Vision (ECCV), pp 135–153, https://doi.org/10.1007/978-3-319-46478-7_9

Wang G, Zhang T, Cheng J, et al (2019a) Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment. In: The IEEE International Conference on Computer Vision (ICCV), pp 3622–3631, https://doi.org/10.1109/ICCV.2019.00372

Wang GA, Yang T, Cheng J, et al (2020) Cross-modality paired-images generation for rgb-infrared person re-identification. In: The AAAI Conference on Artificial Intelligence (AAAI), pp 12,144–12,151, https://doi.org/10.1609/aaai.v34i07.6894

Wang Z, Wang Z, Zheng Y, et al (2019b) Learning to reduce dual-level discrepancy for infrared-visible person re-identification. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 618–626, https://doi.org/10.1109/CVPR.2019.00071

Wei L, Zhang S, Gao W, et al (2018) Person transfer gan to bridge domain gap for person re-identification. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 79–88, https://doi.org/10.1109/CVPR.2018.00016

Wei X, Li D, Hong X, et al (2020a) Co-attentive lifting for infrared-visible person re-identification. The ACM International Conference on Multimedia (ACMMM) pp 1028–1031. https://doi.org/10.1145/3394171.3413933

Wei Z, Yang X, Wang N, et al (2020b) Abp: Adaptive body partition model for visible infrared person re-identification. In: IEEE International

Conference on Multimedia and Expo (ICME), pp 1–6, https://doi.org/10.1109/ICME46284.2020.9102974

Wei Z, Yang X, Wang N, et al (2021) Flexible body partition-based adversarial learning for visible infrared person re-identification. IEEE Transactions on Neural Networks and Learning Systems pp 1–12. https://doi.org/10.1109/TNNLS.2021.3059713

Wu A, Zheng W, Yu H, et al (2017) Rgb-infrared cross-modality person re-identification. In: The IEEE International Conference on Computer Vision (ICCV), pp 5390–5399, https://doi.org/10.1109/ICCV.2017.575

Xiang X, Lv N, Yu Z, et al (2019) Cross-modality person re-identification based on dual-path multi-branch network. IEEE Sensors Journal 19(23):11,706–11,713. https://doi.org/10.1109/JSEN.2019.2936916

Xiao T, Li H, Ouyang W, et al (2016) Learning deep feature representations with domain guided dropout for person re-identification. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 1249–1258, https://doi.org/10.1109/CVPR.2016.140

Ye M, Lanb X, Li J, et al (2018a) Hierarchical discriminative learning for visible thermal person re-identification. In: The AAAI Conference on Artificial Intelligence (AAAI), pp 7501–7508, https://doi.org/https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16734

Ye M, Wang Z, Lan X, et al (2018b) Visible thermal person re-identification via dual-constrained top-ranking. In: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), pp 1092–1099, https://doi.org/10.24963/ijcai.2018/152

Ye M, Lan X, Leng Q (2019) Modality-aware collaborative learning for visible thermal person re-identification. In: The ACM International Conference on Multimedia (ACMMM), pp 347–355, https://doi.org/10.1145/3343031.3351043

Ye M, Lan X, Leng Q, et al (2020a) Cross-modality person re-identification via modality-aware collaborative ensemble learning. IEEE Transactions on Image Processing 29:9387–9399. https://doi.org/10.1109/TIP.2020.2998275

Ye M, Lan X, Wang Z, et al (2020b) Bi-directional center-constrained top-ranking for visible thermal person re-identification. In: IEEE Transactions on Information Forensics and Security, pp 407–419, https://doi.org/10.1109/TIFS.2019.2921454

Ye M, Shen J, J. Crandall D, et al (2020c) Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In: The European Conference on Computer Vision (ECCV), pp 229–247, https://doi.org/10.1007/978-3-030-58520-4_14

Ye M, Shen J, Lin G, et al (2021a) Deep learning for person re-identification: A survey and outlook. IEEE Transactions on Pattern Analysis and Machine Intelligence pp 1–1. https://doi.org/10.1109/TPAMI.2021.3054775

Ye M, Shen J, Shao L (2021b) Visible-infrared person re-identification via homogeneous augmented tri-modal learning. IEEE Transactions on Information Forensics and Security 16:728–739. https://doi.org/10.1109/TIFS.2020.3001665

Zhang C, Liu H, Guo W, et al (2020) Multi-scale cascading network with compact feature learning for rgb-infrared person re-identification. In: International Conference on Pattern Recognition (ICPR), pp 8679–8686, https://doi.org/10.1109/ICPR48806.2021.9412576

Zhang G, Chen Y, Dai Y, et al (2021a) Reference-aided part-aligned feature disentangling for video person re-identification. In: The IEEE International Conference on Multimedia and Expo (ICME), pp 1–6, https://doi.org/10.1109/ICME51207.2021.9428118

Zhang G, Chen Y, Lin W, et al (2021b) Low resolution information also matters: Leai hing multi-resolution representation for person re-identification. In: Proceedings ofthe International Joint Conference on Ariticial Intelligence (IJCAI), pp 1295–1301, https://doi.org/

10.24963/ijcai.2021/179

Zhang G, Ge Y, Dong Z, et al (2021c) Deep high-resolution representation learning for cross-resolution person re-identification. IEEE Transactions on Image Processing 30:8913–8925. https://doi.org/10.1109/TIP.2021.3120054

Zhang G, Yang J, Zheng Y, et al (2021d) Hybrid-attention guided network with multiple resolution features for person re-identification. Information Sciences (INS) 578:525–538. https://doi.org/10.1016/j.ins.2021.07.058

Zhang Q, Lai J, Xie X (2021e) Learning modal-invariant angular metric by cyclic projection network for vis-nir person re-identification. IEEE Transactions on Image Processing 30:8019–8033. https://doi.org/10.1109/TIP.2021.3112035

Zhang X, Luo H, Fan X, et al (2017) Alignedreid: Surpassing human-level performance in person re-identification. arXiv preprint arXiv:171108184

Zhao Y, Lin J, Xuan Q, et al (2019) Hpiln: a feature learning framework for cross-modality person re-identification. IET Image Processing 13(14):2897–2904. https://doi.org/10.1049/iet-ipr.2019.0699

Zheng L, Yang Y, Hauptmann AG (2016) Person re-identification: Past, present and future. arXiv preprint arXiv:161002984

Zhu Y, Yang Z, Wang L, et al (2020) Hetero-center loss for cross-modality person re-identification. Neurocomputing 386:97–109. https://doi.org/10.1016/j.neucom.2019.12.100