The Visual Computer (2023) 39:4117–4132
https://doi.org/10.1007/s00371-022-02579-y

ORIGINAL ARTICLE

Cross-modality collaborative learning identified pedestrian

Xiongjun Wen[1] · Xin Feng[1] · Ping Li[1] · Wenfang Chen[1]

Accepted: 6 June 2022 / Published online: 26 September 2022

Abstract
Cross-modal pedestrian re-identification is a key technology to realize all-weather intelligent video surveillance system. This technology is designed to match the visible light image and infrared image of a pedestrian with a specific identity in a non-overlapping camera scene, so it faces huge intra-class changes and modal differences. Existing methods are difficult to solve these two difficulties, which is largely due to the lack of effective mining of feature discrimination and the full use of multi-source heterogeneous information. In view of the above shortcomings, a refined multi-source feature collaborative network is designed by the collaborative learning method, and multiple complementary features are extracted for information fusion, the learning ability of the network is improved. Multi-scale and multi-level features are extracted from the backbone convolutional network, and the refined feature collaborative learning is realized; the discriminative ability of features is enhanced to deal with intra-class changes. A modal sharing and unique feature collaboration module and a cross-modal human semantic self-supervision module are designed to achieve the purpose of multi-source feature collaborative learning, so as to improve the utilization of multi-source heterogeneous image information, and then modal differences are resolved. The validity and advancement of this method are verified on the SYSU-MM01 and RegDB data sets.

Keywords Pedestrian re-identification · Cross modality · Collaborative learning · Refined features · Multi-source features · Information fusion

1 Introduction

Pedestrian re-recognition is a popular technology in computer vision, and its purpose is to realize pedestrian search in multiple non-overlapping camera scenes [1, 2], it benefits from the vigorous development of pattern recognition and deep learning technology. In recent years, researchers have proposed a series of excellent pedestrian re-recognition methods, and high performance is achieved under ideal simulation conditions [3–5]. However, most of the current methods focus on the images generated by the visible light camera, and in practical applications, the visible light camera can only meet the needs of some scenes. Under night conditions, visible light cameras cannot accurately describe the appearance of pedestrians. Therefore, in order to better meet the conditions of night monitoring, infrared cameras that can image according to temperature become the first choice for night monitoring, and together with the visible light cameras for daytime monitoring, they form an all-weather closed-loop monitoring. To realize such an all-weather intelligent video surveillance system, the main problem is how to match the pedestrian image in the visible light mode with the pedestrian image in the infrared mode, that is, cross-modal pedestrian re-identification.

Cross-modal pedestrian re-recognition is a multi-source fine-grained image retrieval task. The pedestrian images of the two modalities are shown in Fig. 1. What this task needs to match is the image of two different modes of infrared and visible light under a long time span of day and night, so it is more difficult to realize than the traditional single-mode pedestrian re-recognition. These difficulties are mainly reflected in two aspects: (1) Intra-class changes: First, cross-modal pedestrian re-recognition tasks also face intra-class changes in images of pedestrians with the same identity which are caused by factors such as illumination, occlusion, posture, and viewing angle in a single mode. Secondly, the amount of information which is reflected in the infrared image and the visible light image is not equal, and it is very likely that the intra-class change is greater than the inter-class change.

✉ Xin Feng
2801597781@qq.com

[1] School of Information and Mechanical Engineering, Hunan International Economics University, Changsha 410205, Hunan, China

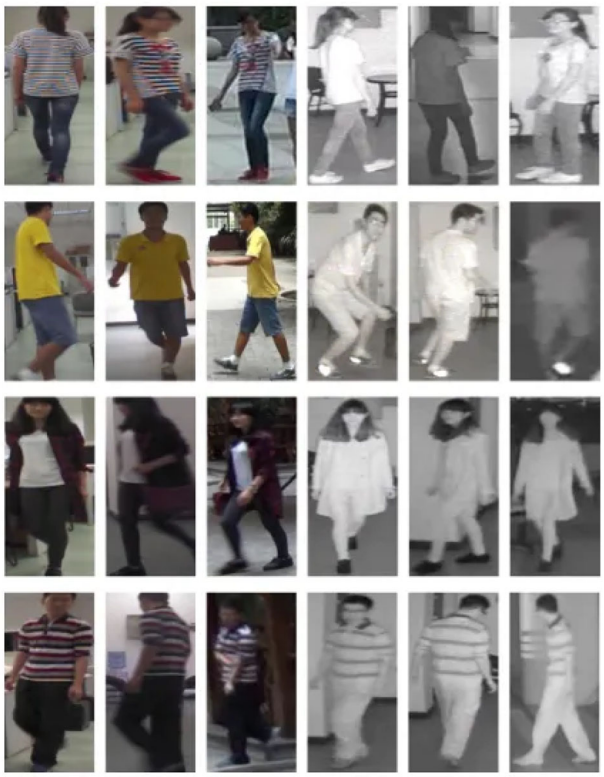


Fig. 1 Examples of images to be processed for cross-modal pedestrian re-recognition. Samples of each row belong to the same person. The first three column images are captured by RGB cameras, while the last three column images are captured by infrared cameras

(2) Modal difference: The problem of cross-modal pedestrian recognition is the mutual retrieval of two heterogeneous images. Feature alignment is the basis for correct image matching. However, due to the different imaging principles of the two images, the distribution of the two images in the feature space is quite different. Therefore, the cross-modal pedestrian re-identification needs to overcome an additional challenge that is the difference between modalities.

For the problem of intra-class change, most of the work often uses the overall characteristics of the image as the final pedestrian representation, and there are also literatures mentioning the strategy of dividing blocks in the horizontal direction. However, only considering the overall features or the local features of a certain scale is one-sided. In addition, the existing work only considers the high-level features which are extracted from the deepest layer of the feature extractor network, and does not consider the low-level features which are obtained from the shallow network. Low-level features can reflect the detailed information of the image, and are also of great significance to the identification of pedestrians. Therefore, to obtain a more discriminative cross-modal pedestrian re-recognition model, a multi-scale and multi-level refined feature extraction strategy can be considered.

Regarding the problem of modal differences, most of the current researches often use shared network parameters, the features of the two images are mapped to the same feature space, and the shared modal feature is used as the final pedestrian representation. However, the features of two modal images can be divided into modal shared and unique features. If only the shared features are considered and the unique features are directly discarded, all the information contained in the image will not be fully utilized. Some studies have found that with the idea of modal conversion, the recognition rate is significantly better than traditional methods. The GAN method can effectively use methods such as style transfer to realize the conversion between the two modalities, and effectively alleviate the difference between the modalities. However, although GAN network can improve task performance to a certain extent, these methods destroy the original spatial structure information and introduce additional noise in the process of reconstructing images or generating features. At the same time, GAN is used, the large amount of calculation and the difficulty of training that are difficult to converge cannot be ignored. Therefore, in dealing with the difference between modalities, it is necessary to fully consider the feature complementarity between the source images of the two modalities of the same person, the utilization of heterogeneous information is improved. Try to achieve neither loss of information nor increase in noise in the process of narrowing the modal difference.

In cross-modal pedestrian re-identification, the various features which are extracted from the image of the same pedestrian have different distributions, but they collectively reflect the identity information of the pedestrian. In this way, the complementarity between various features can be utilized with the help of collaborative learning methods, and the learning ability of the network can be improved through information fusion.

2 Related work

2.1 Monomodal pedestrian re-identification

Single-modal pedestrian re-recognition refers to pedestrian re-recognition considering only visible light modes, which means to solve the problem of matching pedestrian images between non-overlapping visible light cameras [6, 7]. The key challenges of this technology are mainly the intra-class changes in the image of pedestrians with the same identity which is caused by different camera angles, pedestrian posture changes, light intensity [8–13]. The existing single-modal pedestrian re-identification methods can be roughly divided into representation learning methods and metric

learning methods. Representation learning methods mainly use pedestrian identification tags for discriminative feature representation learning [14]. The purpose of metric learning methods is usually to learn the distance between different sample features, and then achieve the effect of increasing the difference between classes and reducing the difference within classes [15]. Early research often used anthropometric data, space and time data, kinematics data, dynamics data and video stream data, etc., specific methods are adopted to describe pedestrian characteristics [16]. Recently, with the help of deep convolutional neural networks, the work of monomodal pedestrian re-recognition has achieved excellent results [9], and even surpassed the recognition level of humans on some, it is widely used public data sets [3, 17]. However, the existing single-modal pedestrian re-recognition method only deals with the pedestrian images which are collected by the visible light camera under good daylight conditions. It is often not well applied in the task of cross-modal pedestrian re-recognition at night [18]. This limits the applicability of this technology to actual all-weather monitoring scenarios.

2.2 Cross-modal pedestrian re-identification

Cross-modal pedestrian re-recognition needs to solve the matching problem between pedestrian images from different imaging sources. The cross-modal pedestrian re-recognition in this paper is the re-recognition of pedestrians between visible light images and infrared images [19–21]. Wu et al. released a large-scale cross-modal pedestrian re-recognition data set SYSU-MM01 [18], analyzed three different network structures and proposed a Deep Zero-padding method. Nguyen et al. published another related data set RegDB [22]. Ye et al. designed a dual-stream network to learn multi-modal shared features [23], while dual-constrained Top-Ranking loss is used to deal with inter-modal and intra-modal changes. In addition, Generative Adversarial Networks (GAN) is used to realize cross-modal pedestrian re-recognition [24], achieving better performance than before, and also providing new ideas for subsequent research work [25, 26]. Later, Zhu et al. considered the local characteristics of the human body in the cross-modal pedestrian re-recognition [27], and introduced the heterogeneous center loss, which greatly improved the recognition accuracy. This paper fully considers the intra-class changes and modal differences between the same identity pedestrian images, and proposes a refined multi-source feature collaborative network. Refined feature collaborative learning methods are used to enhance feature discrimination ability to cope with changes within the class. The multi-source feature collaborative learning method is used to improve the utilization rate of heterogeneous information, modal differences are resolved. Moreover, the effectiveness of this method is verified on the SYSU-MM01 and RegDB data sets.

2.3 Collaborative learning

In order to solve the classification problem, the method theory of collaborative learning is introduced [28]. Collaborative learning refers to training multiple feature learners of the same network on the same batch of training data. The information complementarity between multiple features is used for collaborative fusion, and the generalization ability of the model and the robustness to label noise are improved without increasing the cost of reasoning, so that the network can achieve a better learning effect. Collaborative learning has the advantages of auxiliary training [29], multi-task learning [30, 31] and knowledge distillation [32], but it does not require too many additional training networks and can achieve end-to-end training. It is worth exploring method ideas. For the task of cross-modal pedestrian re-recognition, this paper considers the discriminative ability of multi-scale and multi-level features of deep convolutional neural networks, as well as the information complementarity of multi-source heterogeneous image data, and a collaborative learning method is proposed for refined features and multi-source features.

3 Materials and methods

A refined multi-source feature collaborative network is designed, and its overall network architecture is shown in Fig. 2. For the two modal images of visible light and infrared, how to enhance the feature discrimination ability and improve the utilization of heterogeneous information, and then effectively overcome the two major problems of intra-class variation and modal difference, is the research purpose of the method in this paper. Two parallel ResNet50 in the backbone network form a dual-branch network [33], which are used as feature extractors for visible light and infrared images, respectively. The first stages of the network (Stage1 to Stage4) are used to extract the unique features of each modal, and in the later stages, the shared network parameters are used to extract cross-modal common features. In particular, the designed network in this paper includes a refined feature collaborative learning module (multi-scale feature collaboration and multi-level feature collaboration), and a multi-source feature collaborative learning module (modal sharing and unique feature collaboration and human semantic self-supervision).

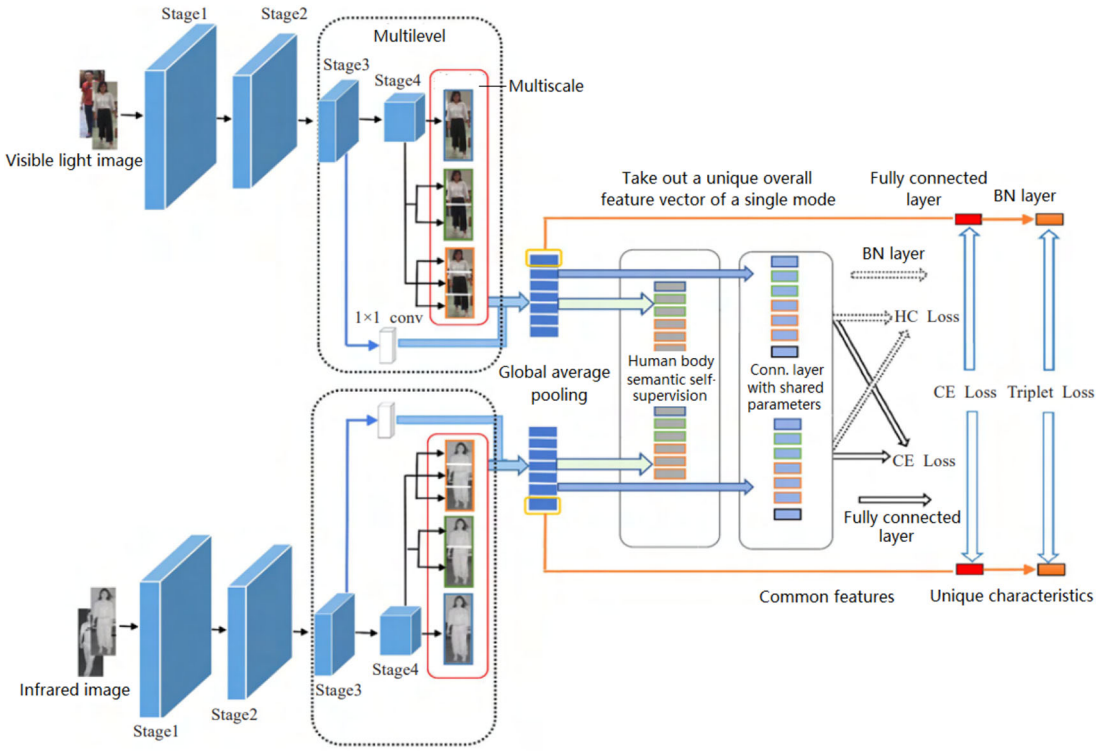


Fig. 2 The overall network architecture for the refined multi-source feature collaborative network

3.1 Collaborative learning of refined features

3.1.1 Multi-scale feature collaboration

Most cross-modal pedestrian re-recognition work is to extract the overall characteristics of the image as the final pedestrian representation [18, 34, 35]. However, because some pedestrians with different identities have small appearance differences, or are affected by occlusion and other noises between cross-modal images, only the overall characteristics are often not effective in distinguishing pedestrian identities. Recently, researchers have proved the effectiveness of using image horizontal block to obtain local features in single-modal and cross-modal pedestrian re-recognition tasks [27]. The local features of different positions will pay attention to different human details, and the details are more distinguishable, so that the model can distinguish different pedestrian identities. However, due to diversified pedestrian posture changes, camera distances and angles, etc., it is sometimes difficult to learn alignment and robust local features with the water scoring method. Therefore, it is not thoughtful to use overall features alone or local features at a specific scale.

The respective advantages and disadvantages of global features and local features are synthesized, a multi-scale feature collaborative learning strategy is proposed in this paper, horizontal multi-scale segmentation is used to deal with cross-modal pedestrian features. For the feature map which is obtained in the fourth stage of ResNet50, a multi-scale block pooling method is used to obtain the multi-scale feature vector of the pedestrian image. In order to obtain partial information of pedestrians at a suitable scale without increasing the amount of calculation, a reasonable block scale needs to be determined. According to the cognition of human joint structure and dressing habits, as well as experimental verification, this paper chooses three block methods: overall, one divided into two, and one divided into three, so that the multi-scale feature collaborative learning method can be used to obtain more discriminative pedestrian information. The work of this paper is the first to comprehensively consider the global and local features at multiple scales, the problem of cross-modal pedestrian re-recognition is solved.

3.1.2 Multi-level feature collaboration

After the feature extractor is performed on the pedestrian image, the features of the corresponding level can be learned from the shallow to the deep stages of the convolutional neural network. In order to intuitively reflect the differences in the features of each layer in the convolutional neural network, ResNet50 is used as the feature extractor, the heat map of the feature distribution at each stage of the network can be obtained as Fig. 3. In the heat map, the distribution of different colors represents the distribution of feature saliency, red represents the most prominent area of the feature, and

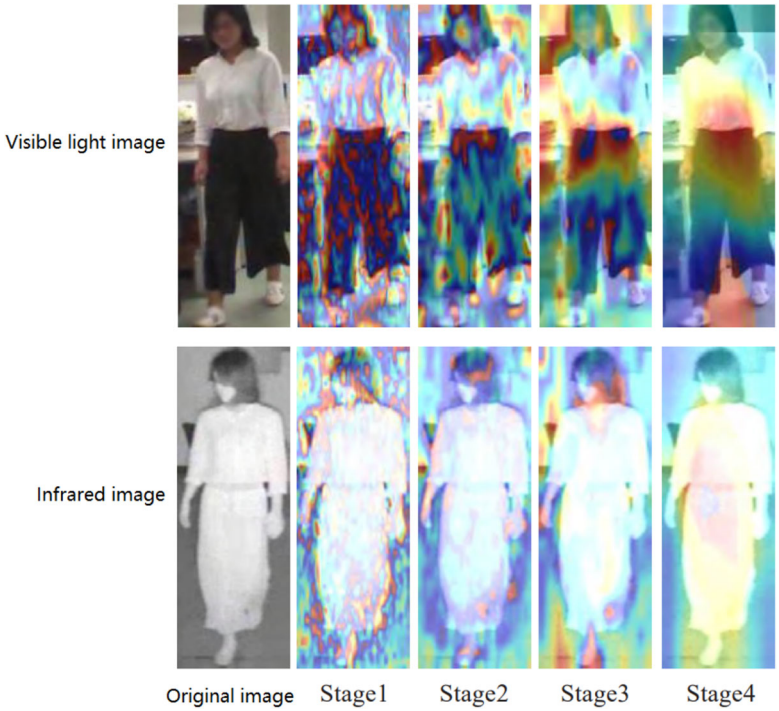


Fig. 3 In ResNet50, each level feature heat distribution map of the pedestrian image

blue represents the most scattered area of the feature. The visual method Grad-CAM (Gradient-weighted Class Activation Mapping) is used for the output of the neural network, which intuitively displays the features learned by the convolutional neural network and helps to understand the working principle and decision-making process of the neural network. Grad-CAM is to use the gradient of any target concept (such as the logits of a class in the classification category, or even the output in the caption task), flow into the final convolutional layer, and generate a rough localization map to highlight the use of images in the image important areas for forecasting.

It is found from Fig. 3 that as the network continues to deepen, the convolutional layers at different stages have noticed changes in the interest areas in the learning process. As shown in Stage1, the first-stage convolutional layer has the most distracted attention, mainly extracting detailed features from the entire pedestrian picture; while in Stage4, the fourth-level convolutional layer focuses on discriminatory features, which mainly extracts key semantic information. Therefore, from the bottom to the top of the convolutional neural network, attention is becoming more and more concentrated, and the extracted information is shifted from scattered spatial structure information to concentrated semantic information.

The current popular pedestrian re-recognition models usually use the deep features of convolutional neural networks to identify pedestrians. However, when learning deep features, due to a large number of filling and merging operations during the training phase, some important spatial information that originally existed in the shallow features, such as shape and texture, will be lost. In addition, infrared images contain less information, which results in a huge difference in the semantic expression capabilities of the two modal images with the same identity. Therefore, it is not appropriate to only use deep features to realize cross-modal pedestrian re-recognition. It is necessary to use the features extracted from the shallow network to supplement the discrimination. A multi-level feature collaborative learning method is proposed in this paper. In order to avoid adding a large amount of calculation and feature dimensions, only the feature map which is obtained in Stage 3 is considered for the shallow information, and the number of channels of this feature map is increased from 1 024 to 2 048 in using 1×1 convolution, and then the shallow features are cascaded with the deep features of Stage4, and they are sent to the following network. Image features of different granularities can be effectively expressed by such multi-level feature coordination strategy, thereby a more discriminative representation of pedestrians is obtained.

Feature Pyramid Network (FPN) is used in here, it is similar to the combination of multi-scale feature fusion and multi-scale prediction, as shown in Fig. 4. Through upsampling combined with lateral connections, the semantic information of high layers is gradually propagated to lower layers. The specific method is to upsample the features of the higher layers by a factor of 2, use 1×1 convolution to change the number of channels on the features of the lower layers, and then add the results of the two. In this way, the feature

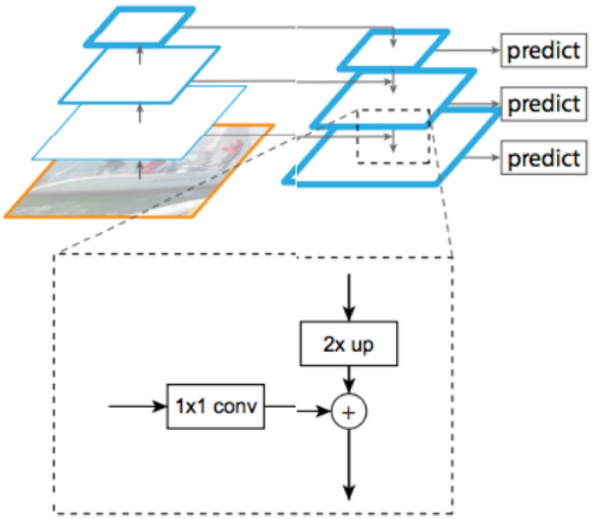


Fig. 4 Feature Pyramid Network (FPN)

maps of each layer are equivalent to merging features of different resolutions, so as to detect objects of corresponding resolutions, so as to ensure that each layer has appropriate resolution and strong semantic features, which can not only detect Small targets, but also ensure strong semantic information for classification.

The receptive field on each scale of FPN is different, and the size corresponding to the original image is different, that is, it can process targets of different scales. After a series of convolutions, the feature map is obtained. Through upsampling, it is restored step by step. In the case of ensuring that the high-level semantic information is not lost, the size of the feature map is also enlarged. Then a large-size feature map is used. To detect small targets. So as to solve the problem that small targets are difficult to detect. In addition, this method consumes little time and computation on the basis of the original network.

3.2 Multi-source feature collaborative learning

3.2.1 Modal sharing and unique feature synergy

The purpose of cross-modal pedestrian re-recognition is to realize the mutual retrieval of two modal images. Due to the differences in different modalities, the description and use of pedestrian representation is a very challenging task in the process of feature learning of cross-modal images. In order to solve this problem, researchers usually use shared network parameters to obtain the common features of the two images as the final pedestrian representation. However, the images of two different modalities of a person contain both modal characteristics and unique modal characteristics. The common feature of the modal can be represented by the intersection of the two sets. If only the common features are considered and the unique features are ignored, it means that the image information is not fully utilized. A cross modality Shared-Specific Transfer Network (cm-SSTN) is proposed [36], which fully considers the shared features between modalities and the unique features within modalities. Their work has achieved the best recognition effect currently, which also verifies the complementary effects of shared features and unique features. However, cm-SSTN also has shortcomings such as complex model and large amount of calculation.

Based on the heterogeneous information complementarity of cross-modal image features, a simple and effective method is proposed for collaborative learning of modal shared and unique features. On the basis of the cross-modal two-branch network, a fully connected layer with parameter sharing is used to extract the common features of the modal, and a fully connected layer with no shared parameters is used to extract the unique characteristics of the modal. Then, in the supervised learning, the modal shared feature and the modal unique feature are trained separately to achieve the effect of heterogeneous complementarity and improve the utilization of image information. Fully connected layers (FC) play the role of "classifier" in the entire convolutional neural network. If the operations such as convolutional layer, pooling layer and activation function layer are to map the original data to the hidden layer feature space, the fully connected layer plays the role of mapping the learned “distributed feature representation” to the sample label space.

3.2.2 Human body semantic self-supervision

Information interaction between modalities is an effective means to reduce modal differences. The existing work usually uses GAN network to realize the style transfer or feature transfer of the image, the information interaction is realized between the modalities. However, the GAN network may introduce new noise based on the idea of generation, and it will face problems such as difficulty in convergence during training. Therefore, how to learn knowledge that is not constrained by modal characteristics without introducing noise and easy to train, and then the information interaction is realized between modals, is an idea worth exploring.

A human body semantic self-supervised module is proposed in this paper, which aims to use the semantic information of human body structure as prior knowledge, and send two modal images to a self-supervised learning network of shared parameters without using identity tags, and learn the basic characteristics of the human body that are not restricted by modalities and identity. That is to say, from an objective perspective, no matter which identity the pedestrian image belongs to and which modality it comes from, the relative position of each body part of the human body in a pedestrian image is determined. In other words, from top to bottom, each person’s image is a semantic structure such as

head, shoulders, chest, abdomen, legs, and feet. This semantic information is an obvious difference between pedestrian images and other natural images, and it is also an important prior knowledge. It happens to use this prior knowledge to design a self-supervised module of human body semantics with cross-modal parameter sharing.

The human body parts in each image of different modalities are divided into small blocks and the order is shuffled, and then the blocks are reordered under the supervision of each block position label, the original order is obtained. In this way, the relative positional relationship of various parts of the human body can be used to learn basic human body information that has nothing to do with the modal source of the image, so as to achieve the effect of reducing modal differences. Experiments prove that this simple operation can get better results. Moreover, the self-supervised module uses the multi-scale feature block of a single pedestrian image in the above multi-scale collaborative learning method, it does not introduce a lot of calculation.

3.3 Loss function design

Cross-modal pedestrian re-recognition networks often use Cross Entropy Loss (CE Loss) and Triplet Loss to supervise learning features [34, 35]. Cross-entropy loss is used to classify pedestrian identities, and the triplet loss is used to reduce the intra-class distances and increase the inter-class distances. Later, a Hetero Center Loss (HC Loss) was proposed [27]. The purpose of designing this loss function is to reduce the difference between similar samples of different modalities. Heterogeneous center loss and cross-entropy loss are used in this work, the better results are achieved.

In the method in this paper, a mix-Modality Triplet Loss is introduced [37], and it is used in combination with cross-entropy loss and heterocenter loss. It is proved that the cross-entropy loss and the triplet loss function in the same feature space will cause convergence difficulties [38]. In the same way, there is also such a problem between the cross-entropy loss and the heterocenter loss. Therefore, a batch normalization layer (BN Layer) and a fully connected layer (FC Layer) are used to map feature vectors to two feature spaces, the conflicts are resolved.

The size of each batch of input images is denoted as N, then $N = 2 \times P \times K$, which means that there are P pedestrian identities in N pictures in each batch, and each identity has K visible light images and K infrared image. For modal shared features, the pedestrian identity information of each picture is used as the supervision label [27], and the combined effect of cross-entropy loss and heterogeneous center loss is used to learn each feature block. The calculation method of cross entropy loss on each feature block is formula (1):

$$L_{sh-CE} = -\sum_{i=1}^{2\times P\times K} p(x_i) lb(q(x_i)) \tag{1}$$

wherein, x_i refers to a certain feature block of the i-th image, and $p(x_i)$ refers to the expected output, that is, the real label. $q(x_i)$ is the predicted label obtained after the extracted feature vector of each feature block in the network passes through the *Softmax* layer. The heterogeneous center loss is calculated for each feature block, it is in formula (2):

$$L_{sh-HC} = \sum_{p=1}^{P} \left|\left|c_{p,1} - c_{p,2}\right|\right|_2^2 \tag{2}$$

Here, $c_{p,1} = \frac{1}{K}\sum_{j=1}^{K} x_{p,1j}$ and $c_{p,2} = \frac{1}{K}\sum_{j=1}^{K} x_{p,2j}$ are heterogeneous centers, which are the sample center points of K samples in the visible light and infrared modalities of the image representing the identity of the p-th pedestrian. Therefore, $\left|\left|c_{p,1} - c_{p,2}\right|\right|_2^2$ refers to the distance between the sample centers of the p-th pedestrian identity in the two modalities. Therefore, the calculation method of the loss function on the common features is as formula (3):

$$L_{SH} = \sum_{f=1}^{7} (L_{sh-CE,f} + \lambda L_{sh-HC,f}) \tag{3}$$

wherein, λ is a weight parameter that balances the cross-entropy loss and the heterogeneous center loss. f from 1 to 7 refers to calculating the total loss of 7 feature blocks. The 7 feature blocks are one feature block in the shallow layer and 6 multi-scale feature blocks in the deep layer.

For the unique characteristics of the modal, the cross-entropy loss is first used to identify the identity of each sample, as is shown in Eq. (4):

$$L_{sp-CE} = -\sum_{i=1}^{2\times P\times K} p(g_i) lb(q(g_i)) \tag{4}$$

Here, g_i is used to denote the overall feature vector obtained from the single-mode branch. In addition, the triplet loss is used to reduce the intra-class difference and increase the inter-class difference. The triplet loss calculation requires three input images, which are a fixed image (Anchor) a, a positive sample image (Positive) p, and a negative sample image (Negative) n. Images a and p are positive sample pairs, and images a and n are negative sample pairs. Taking into account that other parts of the network have played a role in reducing the modal difference, a mixed modal triplet loss function is used here, that is, the two modal sample features are placed in the same set for triple sampling. Then the number of pictures

in a batch is $2 \times P \times K$, the set of all pictures in a batch is *batch*, the positive sample set of the fixed image a is A, and the negative sample set is B. Then the calculation method of mixed-mode triplet loss is as formula (5):

$$L_{sp-m\text{Tri}} = \frac{1}{2PK} \sum_{a \in \text{batch}} \left[\alpha + \max_{p \in A} d_{a,p} - \min_{n \in B} d_{a,n} \right]_+ \quad (5)$$

The α in the formula refers to the boundary value parameter of the triplet loss. $[.]_+$ means that if the calculation result in the square brackets is less than 0, it is recorded as 0. A and B are both subsets of *batch*. The loss function of the unique characteristics of the modal is the formula (6):

$$L_{SP} = L_{sp-CE} + L_{sp-m\text{Tri}} \quad (6)$$

For the human semantic self-supervised module, the purpose is to reconstruct and sort the disrupted block feature vectors during the self-supervised training process. The specific method is that the position of the six multi-scale feature blocks is labeled, and then label is predicted during the training process, and then the original spatial relative positions of the 6 block feature vectors are learned. The cross-entropy loss function of the predicted feature block label can be used as the loss function of self-supervised learning, and $S_{i,s}$ is the s-th block of the i-th sample, and the loss function of this self-supervised learning module can be calculated as the formula (7):

$$L_{\text{SSL}} = -\sum_{i=1}^{2\times P\times K} \sum_{s=1}^{6} p(S_{i,s}) lb(q(S_{i,s})) \quad (7)$$

In formula (7), $p(S_{i,s})$ is the true position label of each feature block, and $q(S_{i,s})$ is the predicted label of each feature block.

In summary, the total loss function of this refined multi-source feature collaborative network in the end-to-end training process is Eq. (8):

$$L = L_{SH} + L_{SP} + L_{SSL} \quad (8)$$

4 Experiment and performance analysis

4.1 Experimental setup

4.1.1 Data set

There are currently two public data sets (SYSU-MM01 [18] and RegDB [22]), they can be used to evaluate the experimental results of cross-modal pedestrian re-recognition methods. The images in the data set are collected from visible light cameras and infrared (near infrared and far infrared) cameras.

The SYSU-MM01 data set is a large-scale data set which are collected by six different cameras in outdoor and indoor environments, including four visible light cameras and two near-infrared cameras. The data set contains training data for 395 pedestrian identities, including 22,258 visible light images and 11,909 near-infrared images. The test set contains another 95 images of pedestrian identities, as well as two evaluation modes and two test set construction methods. In the two evaluation modes, the query set (Qurey set) is the same, containing 3,803 images which are captured from two infrared cameras. In All-search mode, the Gallery set contains all visible light images captured from all four visible light cameras. In Indoor-search mode, the gallery collection only contains visible light images captured by two indoor visible light cameras. Generally speaking, All-search is more challenging than Indoor-search mode. The two test set construction methods are Single-shot and Multi-shot. The method of the two is to randomly select 1 or 10 pictures of the same pedestrian identity when constructing the gallery set. A detailed description of the evaluation scheme can be found in the literature [17]. The most difficult experimental setting is used, that is, the All-search evaluation mode and the Single-shot test set construction method, 10 tests were performed and the average retrieval performance was recorded.

The RegDB data set is a small-scale data set collected by a dual-mode camera system (a visible light camera and a far-infrared camera). In the RegDB data set, the visible image and infrared image contours are very similar, and cross-modal pedestrian recognition is less difficult. This data set contains a total of 412 pedestrian identities, each of which has 10 visible images and 10 infrared images. According to the evaluation protocol [22], 206 identities (2,060 images) are randomly selected for training, and the remaining 206 identities (2,060 images) are used for testing. The performance of two different retrieval settings, namely visible light image retrieval infrared image (Visible to Thermal) and infrared image retrieval visible light image (Thermal to Visible) are evaluated, and the average accuracy is recorded by randomly dividing the training set and the test set 10 times.

4.1.2 Evaluation index

For the sake of fairness, in this experiment, Cumulative Matching Characteristics (CMC) and mean Average Precision (mAP) are used as evaluation indicators. The Rank accuracy rate in CMC is the probability that the correct cross-modal pedestrian image appears in the first k retrieval results. The mAP index can reflect the average retrieval performance of the method.

Table 1 Comparison of the method in this paper with other methods on the SYSU-MM01 data set (%)

Method	Source	All-search & single-shot			
		Rank1	Rank10	Rank20	mAP
Zero-Padding[18]	ICCV17	14.80	54.12	71.33	15.95
HCML[39]	AAAI18	14.32	53.16	69.17	16.16
cmGAN[24]	IJCAI18	26.97	67.51	80.56	27.8
HSME[40]	AAAI19	20.68	62.74	77.95	23.12
D2RL[25]	CVPR19	28.90	70.60	82.40	29.20
AlignGAN[26]	ICCV19	42.40	85.00	93.70	40.70
HPILN[41]	TIP19	41.36	84.78	94.51	42.95
eBDTR[23]	TIFS20	27.82	67.34	81.34	28.42
Hi-CMD[35]	CVPR20	34.94	77.58	–	35.94
JSIA[26]	AAAI20	38.10	80.70	89.90	36.90
MSR[42]	TIP20	37.35	83.40	93.34	38.11
AGW[43]	ECCV20	47.50	–	–	47.65
XIV [44]	AAAI20	49.92	89.79	95.96	50.73
HAT[45]	TIFS20	55.29	92.14	97.36	53.89
SIM[46]	IJCAI2020	56.93	91.50	96.82	60.88
TSLFN + HC[27]	Neuro20	56.96	91.50	96.82	54.95
cm-SSFT[36]	CVPR2020	61.60	–	–	63.20
Ours		66.24	96.74	99.26	65.40

4.1.3 Experimental design details

In the experiment, the Pytorch framework is used to implement the engineering code, and training and testing are conducted on an NVIDIA GeForce 1080Ti GPU. The size of the pedestrian image in the data set is adjusted to 384 × 128. In the training phase, four pedestrian identities are randomly selected, and then eight visible light images and eight infrared images are randomly selected for each pedestrian identity. Therefore, in each round of training, the batch size is 64. In order to balance the effects of the cross-entropy loss function and the heterogeneous center loss function, the weight of the heterogeneous center loss is set to 0.5 in formula (1) [27]. The boundary value of the triplet loss is set to 0.3. A stochastic gradient descent (SGD) optimizer with momentum of 0.9 is used in the training process, including the training process of the Warm Up Learning Rate strategy which is adopted in the first 10 rounds, the refined multi-source feature collaborative network has been trained for 80 rounds. The learning rate *lr(t)* changes with the training round t, as shown in formula (9):

$$lr(t) = \begin{cases} t, \ t \leq 10 \\ 0.1, \ 10 < t \leq 20 \\ 0.01, \ 20 < t \leq 50 \\ 0.001, \ 50 < t \leq 80 \end{cases} \tag{9}$$

In the training process, the network is optimized by using modal common and modal characteristics. When testing reasoning, only common modal features are used to evaluate the similarity between the query image and the gallery image. The reason is firstly that under the influence of the unique characteristics of the modal, the training is finally completed through end-to-end collaborative learning, and the extracted common characteristics of the modal can effectively describe the image, which is proved in the experiment of this article. Another reason is that the use of shared features alone can speed up the feature inference during the testing process.

4.2 Performance analysis

4.2.1 Comparative analysis with other methods

On the SYSU-MM01 and RegDB data sets, the method in this paper is compared with some popular methods of current cross-modal pedestrian re-recognition tasks under the same experimental setup. These methods include Zero-Padding [18], HCML [39], cmGAN [24], HSME [40], D2RL [25], AlignGAN [26], HPILN [41], eBDTR [23], Hi- CMD [35], JSIA [26], MSR [42], AGW [43], XIV [44], HAT [45], SIM [46], EDFL [47], TSLFN + HC [27] and cm-SSFT [36]. The experimental results are shown in Tables 1 and 2.

In Table 1, the refined multi-source feature collaborative network is similar to TSLFN + HC, but the method in this

Table 2 Comparison of the method in this article and other methods on the RegDB data set

Method	Source	Rank1	Rank10	Rank20	mAP
(a) Visible to thermal (%)					
Zero-Padding[18]	ICCV17	17.75	34.21	44.35	18.90
HCML[39]	AAAI18	24.44	47.53	56.78	20.80
D2RL[25]	CVPR19	43.40	66.10	76.30	44.10
HSME[40]	AAAI19	50.85	73.36	81.66	47.00
AlignGAN[26]	ICCV19	57.90	–	–	53.60
eBDTR[23]	TIFS20	34.62	58.96	68.72	33.46
MSR[42]	TIP20	48.43	70.32	79.95	48.67
JSIA[26]	AAAI20	48.50	–	–	48.90
XIV [44]	AAAI20	62.21	83.13	91.72	60.18
Hi-CMD[35]	CVPR20	70.93	86.39	–	66.04
AGW[[43]	ECCV20	70.05	–	–	66.37
HAT[45]	TIFS20	71.83	87.16	92.16	67.56
cmSSFT[36]	CVPR20	72.30	–	–	72.90
Ours		86.89	93.74	96.31	84.72
(b) Thermal to visible (%)					
Zero-Padding[18]	ICCV17	16.63	34.68	44.25	17.82
HCML[39]	AAAI18	21.70	45.02	55.58	22.24
eBDTR[23]	TIFS20	34.21	58.74	68.64	32.49
HSME [40]	AAAI19	50.15	72.40	81.07	46.16
EDFL[47]	Neuro20	51.89	72.09	81.04	52.13
AlignGAN[26]	ICCV19	56.30	–	–	53.40
Ours		62.88	81.67	87.11	58.34

paper leads 9.28% in Rank1 and 10.45% in mAP. In addition, cm-SSFT is the best of all comparison methods. Although cm-SSFT reached 61.60% and 63.20% in Rank1 and mAP, respectively, the experimental results of this paper showed that Rank1 and mAP were 4.64 and 2.20 percentage points higher than cm-SSFT, respectively. Moreover, cm-SSFT has a more complex network structure, which brings more parameters and calculations.

It is seen from Table 2 that the method in this paper is also competitive on the RegDB data set, and the recognition accuracy is higher than that on the SYSU-MM01 data set. This is largely because the images of the RegDB data set are collected from dual-mode cameras, and the obtained visible light images are similar to the pedestrians in the infrared images, so the cross-modal intra-class differences are small. In addition, the experimental data in Table 2 show that the recognition effect of visible light image retrieval infrared image mode is higher than that of infrared image retrieval visible light image mode. This is because the pedestrian image in the infrared mode has a small amount of information, and the ability to discriminate the identity of the pedestrian is not strong. This feature is also consistent with the above-mentioned viewpoint.

4.2.2 Visual analysis of retrieval results

In order to visually analyze the re-identification effect of the proposed method in this paper, several samples in the SYSU-MM01 data set were selected for visual analysis of the retrieval results, as shown in Fig. 5. The first three lines in the figure are the results of using infrared images to retrieve visible light images, and the last three lines are the results of using visible light images to retrieve infrared images. The first column in the figure is the search target pedestrian image. The remaining columns are the top 10 pedestrian images in the search results. From left to right, the pictures are sorted according to the similarity which is calculated by the model in descending order. The green box in the figure is the sample that was retrieved correctly, and the red box is the sample that was retrieved incorrectly.

As shown in Fig. 5, although the color of the upper and lower body clothes in the visible light pedestrian image is quite different, there is no obvious difference in the infrared mode. If such a sample needs to be paired correctly, the model needs to pay more attention to the movement, body shape and some details of the pedestrian texture characteristics. It can be seen from the search results that the method in this paper can effectively extract refined pedestrian features.

Fig. 5 In this paper method, the re-identification effect on the SYSU-MM01 data set

It is seen from Fig. 5 that the common features between modalities, such as bag and clothing logos, will still be the key to information matching in the recognition process, and these common features of modalities may be helpful to the judgment of the correct result. Therefore, it is very important to adopt a multi-source feature collaborative learning method to promote information interaction between modalities and extract more discriminative features.

In addition, when the color cannot be used as identification information, the deep network will learn the pedestrian's body shape, posture and other characteristics, these are used as an important basis for distinguishing pedestrians. As shown in the sixth row of Fig. 7, although the first, second, and fourth columns are pedestrians with different identities, they are misjudged as the same person because they all cross their legs. It can be seen that extracting reliable discriminative features is still an important challenge.

4.3 Ablation experiment

In order to verify the effectiveness of the various modules of the refined multi-source feature collaborative network proposed in this paper, an ablation experiment was performed on the network. On the SYSU MM01 data set, the TSLFN + HC [27] method is used as the baseline model, and several modules proposed in this paper are sequentially added to the network, which can clearly quantify and reflect the improvement effect of each module on the task.

It can be seen that the various modules proposed in this paper are helpful for cross-modal pedestrian re-recognition tasks. For each experiment in Table 3, the module design analysis is carried out below.

4.3.1 Multi-scale feature collaboration

As in Experiment 2 in Table 3, the local features of different scales are obtained through several level division strategies, and they are cascaded with the overall features to obtain the multi-scale feature module in this paper. In order to determine the optimal scale of the multi-scale feature collaboration module, several different levels of horizontal block combination strategies are compared. TSLFN + HC is used as the baseline method (Baseline) to analyze the effect of module design. In other words, in this experiment, only the feature level six-division method of the TSLFN + HC method is changed, and the rest of the network structure and experimental settings are unchanged. The combinations are as follows: Scale1 (global feature + 2 horizontal equal blocks), Scale2

Table 3 Comparison of the method in this paper with other methods on the SYSU-MM01 data set (%)

Experiment no.:	TSLFN + HC (Baseline)	Multi-scale feature collaboration	Multi-level feature collaboration	Co-ordination of common and unique characteristics	Human body semantic self-supervision	Performance/%	
						Rank1	mAP
1	√					56.96	54.95
2	√	√				61.23	59.85
3	√	√	√			64.69	62.38
4	√	√	√	√		65.08	63.85
5	√	√	√	√	√	66.24	65.40

(global feature + 2 horizontal equal blocks + 3 horizontal equal blocks), Scale3 (global feature + 2 horizontal equal block feature + 3 horizontal equal block features + 4 horizontal equal block features). As shown in Fig. 6, the best feature is Scale2. Moreover, according to objective cognition, the human body structure level is divided into two or three parts, which can be understood as independent semantic units, so Scale2 is suitable for personal representation.

4.3.2 Multi-level feature collaboration

As in Experiment 3 in Table 3, after the best multi-scale features have been selected, experiments are also carried out to find the best multi-level features. Features are extracted at different levels and different combinations are analyzed. Level2 and Level3 represent different feature maps which are extracted from Stage1 and Stage2 based on the Resnet50 backbone network. The results are shown in Table 4. The best shallow feature is Level3. The features extracted from Level2 will degrade performance in any combination. For example, the performance of Multi-Scale + Level2 is lower than Multi-Scale, and the performance of Multi-Scale + Level2 + Level3 is also lower than Multi-Scale + Level3. It can be seen that the level of feature information extracted by Level2 is too low and does not significantly contribute to semantic classification.

4.3.3 Modal sharing and unique feature synergy

As shown in Experiment 4 in Table 3, after multi-scale and multi-level methods are used to achieve refined feature collaborative learning, a modal shared and unique feature collaborative learning module was designed. The mixed modal triplet loss function is an important part of the proposed collaborative learning module of modal shared and unique features. The reason for using such a loss function is to sample the triples by mixing two modal images in a training batch, so that the inter-modal information exchange can be better realized in the process of metric learning. In order to verify the function of the mixed-mode triplet loss function and its advantages over the single-mode triplet loss, a comparative experiment was carried out on the two under the condition that the other design parts of the network remain unchanged. The results are shown in Fig. 7.

4.3.4 Human body semantic self-supervision

As in Experiment 5 in Table 3, the self-supervised learning module of human semantics improves the performance of cross-modal pedestrian re-recognition tasks. From a logical analysis, the input data of the module is the characteristics of the two modalities, which can achieve the effect of overcoming the difference of modalities. However, this module

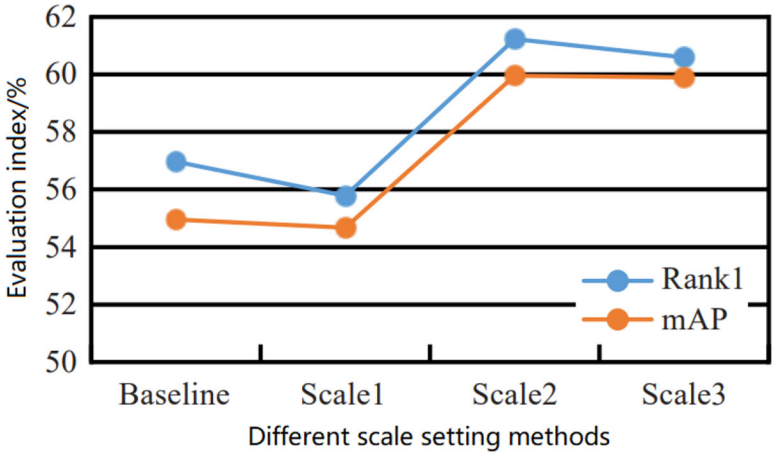


Fig. 6 Examples of images to be processed for cross-modal pedestrian re-recognition

Table 4 Performance analysis of various hierarchical combination methods (%)

Method	Rank1	mAP
Multi-scale (Only use multi-scale feature collaboration)	61.23	59.85
Multi-scale + level2	60.66	59.76
Multi-scale + level3	64.69	62.38
Multi-scale + level2 + level3	63.77	61.58

Table 5 Performance analysis of human body semantic self-supervised module (%)

Method	Rank1	mAP
No self-supervision	65.08	63.85
Monomodal self-supervision	65.23	64.17
Cross-modal self-supervision	66.24	65.40

can also play a role in local feature learning. Therefore, the performance improvement is brought by this module, it is to realize the overcoming of modal differences or the realization of local feature learning.

On the premise that the other designs of the network remain unchanged, the three settings of no self-supervision, single-modal self-supervision, and cross-modal self-supervision are compared, as shown in Table 5. It can be seen that cross-modal human semantic self-supervision not only has the function of local feature learning, but also can achieve the effect of overcoming modal differences.

5 Conclusions and outlook

Visible thermal person re-identification (VT-ReID) is a challenging cross-modality pedestrian retrieval problem due to the large intra-class variations and modality discrepancy

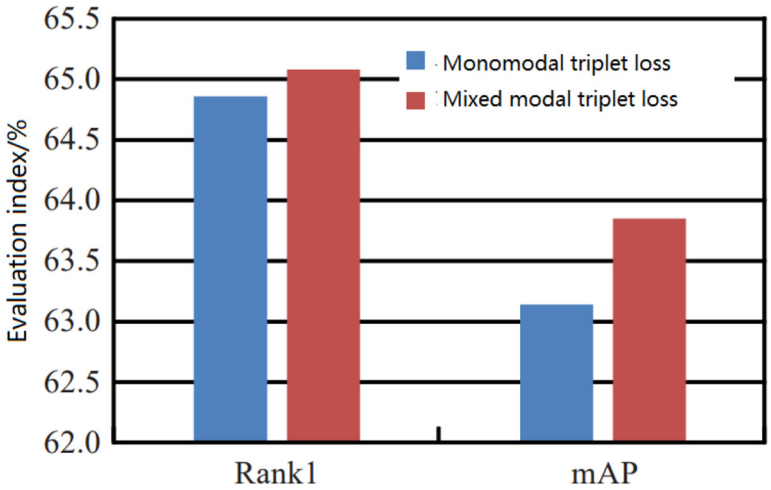


Fig. 7 Performance comparison of triplet loss function under different conditions

across different cameras. In classifier level, both modality-specific and modality-sharable identity classifiers for two modalities are introduced to handle the modality discrepancy [48]. To utilize the complementary information among different classifiers, we propose an ensemble learning scheme to incorporate the modality sharable classifier and the modality specific classifiers. In addition, we introduce a collaborative learning strategy, which regularizes modality-specific identity predictions and the ensemble outputs.

In this paper, comprehensive consideration is given to enhancing the ability of feature discrimination and improving the utilization of multi-source heterogeneous information. Under the guidance of collaborative learning methods, a refined multi-source feature collaborative network is proposed. Multi-scale and multi-level features are used to achieve refined feature collaborative learning, and the purpose of multi-source feature collaborative learning is achieved through modal sharing and unique feature collaboration and human semantic self-supervision. The proposed method in this paper is obviously superior to other methods on two related data sets, and provides a simple and effective idea for further research in this field.

The main contributions of this article are as follows:

(1) In order to enhance the ability to distinguish features, in this paper, a collaborative learning method is proposed for refined features, that is, when designing a convolutional neural network for feature extraction, multi-scale and multi-level pedestrian features are comprehensively considered. Experiments show that collaborative learning of refined features is a simple and effective method, the ability of feature discrimination is enhanced.
(2) In order to improve the utilization of multi-source heterogeneous information, a multi-source feature collaborative learning method is proposed this paper. First, in view of the heterogeneous information complementarity of visible light images and infrared images, the dual-stream network is used to extract the common and unique features of cross-modal images for collaborative learning. Secondly, a priori identification of the relative positional relationship of various parts of the human body is used as an auxiliary task, and a human body semantic self-supervision method is proposed. Finally, the purpose of multi-source feature collaborative learning is achieved under the joint supervision of multiple targeted loss functions.
(3) Sufficient experiments were conducted on the relevant data set of cross-modal pedestrian re-identification. It is verified that the performance of the refined multi-source feature collaborative network in this paper is better than the current best related work, and it has higher reliability and advancement.

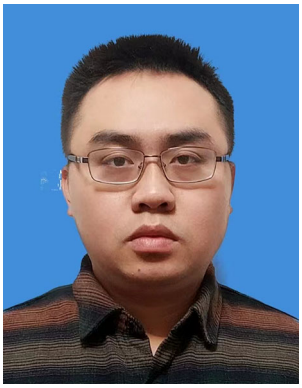
Acknowledgements This work was supported by First-class course in Hunan Province project ([2021] 322, No.167); Hunan University Student Innovation and Entrepreneurship Training Program: ([2022] 174, No. 4531 [2021] 197. No. 3281): Teaching Reform Research Proiec: Xiangwalingvuan [2022] No. 64.

References

1. Song, L.L., Li, B., Zhao, J.Y., et al.: Normality resampling of improved metric learning method for person re-identification. Comput. Eng. Appl. **56**(8), 158–165 (2020). https://doi.org/10.3778/j.issn.1002-8331.1904-0235
2. Fan, C.X., Chen, Y.J., Cao, L., et al.: Person re-identification based on visual perceptual model. Comput. Eng. Appl. **52**(6), 156–161 (2016). https://doi.org/10.3778/j.issn.1002-8331.1504-0307
3. Sun, Y., Zheng, L., Yang, Y., et al.: 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)
4. Wang, G., Yuan, Y., Chen, X, et al.: Learning discriminative features with multiple granularities for person re-identification. In: Proceedings of the 26th ACM International Conference on Multimedia, pp. 274–282 (2018)
5. He, L., Liao, X., Liu, W., et al.: Fastreid: a pytorch toolbox for general instance re-identification. arXiv:2006.02631arXiv:2006.02631 (2020).
6. Zheng, L., Yang, Y., Hauptmann, A. G.: Person re-identification: past, present and future. arXiv:1610.02984 (2016).
7. Zhu, X., Wu, B., Huang, D., et al.: Fast open-world person re-identification. IEEE Trans. Image Process. **27**(5), 2286–2300 (2017)
8. Bai, S., Tang, P., Torr, P. H. S., et al.: Re-ranking via metric fusion for object retrieval and person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 740–749 (2019)
9. Dai, J., Zhang, P., Wang, D., et al.: Video person re-identification by temporal residual learning. IEEE Trans. Image Process. **28**(3), 1366–1377 (2018)
10. Hou, R., Ma, B., Chang, H. et al. Interaction-and-aggregation network for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9317–9326 (2019)
11. Song, C., Huang, Y., Ouyang, W., et al.: Mask- guided contrastive attention model for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1179–1188 (2018)
12. Sun, Y., Xu, Q., Li, Y., et al.: Perceive where to focus: learning visibility- aware part-level features for partial person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 393–402 (2019)
13. Yang, W., Huang, H., Zhang, Z., et al.: Towards rich feature discovery with class activation maps augmentation for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1389–1398 (2019)
14. Liao, S., Hu, Y., Zhu, X., et al.: Person re-identification by local maximal occurrence representation and metric learning. In: Proceedings of the IEEE Conference on Com puter Vision and Pattern Recognition, pp. 2197–2206 (2015)
15. Liao, S., Li, S. Z.: Efficient psd constrained asymmetric metric learning for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3685–3693 (2015)
16. Ben, X.Y., Xu, S., Wang, K.J.: Review on pedestrian gait feature expression and recognition. Pattern Recognit. Artif. Intell.

25(1), 71–81 (2012). https://doi.org/10.3969/j.issn.1003-6059.2012.01.010
17. Zhang, X., Luo, H., Fan, X., et al.: AlignedReID: surpassing human-level performance in person re-identification. arXiv:1711.08184, (2017)
18. Wu, A., Zheng, W. S., Yu, H. X., et al.: Rgb-infrared cross modality person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5380–5389 (2017)
19. Zheng, L., Bie, Z., Sun, Y., et al.: Mars: a video benchmark for large-scale person re-identification. In: European Conference on Computer Vision, pp. 868–884. Springer, Cham (2016).
20. Wu, Y., Lin, Y., Dong, X., et al.: Progressive learning for person re-identification with one example. IEEE Trans. Image Process. **28**(6), 2872–2881 (2019)
21. Chen, D., Li, H., Liu, X., et al.: Improving deep visual rep resentation for person re-identification by global and local image-language association. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 54–70 (2018)
22. Nguyen, D.T., Hong, H.G., Kim, K.W., et al.: Person recognition system based on a combination of body images from visible light and thermal cameras. Sensors **17**(3), 605 (2017)
23. Ye, M., Lan, X., Wang, Z., et al.: Bi-directional center-constrained top-ranking for visible thermal person re-identification. IEEE Trans. Inf. Forensics Secur. **15**, 407–419 (2019)
24. Dai, P., Ji, R., Wang, H., et al.: Cross-modality person re-identification with generative adversarial training. In: 27th International Joint Conference on Artificial Intelligence, (2018)
25. Wang, Z. X., Wang, Z., Zheng, Y., et al.: Learning to reduce dual-level discrepancy for infrared- visible person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 618–626
26. Wang, G., Zhang, T., Cheng, J., et al.: Rgb- infrared cross- modality person re- identification via joint pixel and feature alignment. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3623–3632 (2019)
27. Zhu, Y., Yang, Z., Wang, L., et al.: Hetero- center loss for cross-modality person re-identification. Neurocomputing **386**, 97–109 (2020)
28. Song, G., Chai, W.: Collaborative learning for deep neural networks. Adv. Neural Inf. Process. Syst. **31**, 1832–1841 (2018)
29. Szegedy, C., Liu, W., Jia, Y., et al.: Going deeper with convolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 7, pp. 1–9 (2015)
30. Caruana, R.: Multitask learning. Mach. Learn. **28**(1), 41–75 (1997)
31. Baxter, J.: Learning internal representations. In: Proceedings of the Eighth Annual Conference on Computational Learning Theory, pp. 311–320 (1995)
32. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowl edge in a neural network. arXiv:1503.02531, (2015)
33. He, K., Zhang, X., Ren, S., et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
34. Ye, M., Lan, X., Li, J., et al.: Hierarchical discriminative learning for visible thermal person re-identification. In: Thirty- Second AAAI Conference on Artificial Intelligence, (2018)
35. Choi, S., Lee, S., Kim, Y., et al.: 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)
36. Lu, Y., Wu, Y., Liu, B., et al.: 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)
37. Wang, J., Jiao, S., Li, Y., et al.: Two-stage metric learn ing for cross- modality person re-identification. In: Proceedings of the 5th International Conference on Multi media and Image Processing, pp. 28–32 (2020)
38. Luo, H., Jiang, W., Gu, Y., et al.: A strong baseline and batch normalization neck for deep person re-identification. IEEE Trans. Multimed. **22**(10), 2597–2609 (2019)
39. Ye, M., Lan, X., Li, J., et al.: Hierarchical discriminative learning for visible thermal person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence, (2018)
40. Hao, Y., Wang, N., Li, J., et al.: HSME: hypersphere manifold embedding for visible thermal person re-identification. Proc. AAAI Conf. Artif. Intell. **33**, 8385–8392 (2019)
41. Zhao, Y.B., Lin, J.W., Xuan, Q., et al.: HPILN: a feature learning framework for cross-modality person re-identification. IET Image Proc. **13**(14), 2897–2904 (2019)
42. Feng, Z., Lai, J., Xie, X.: Learning modality-specific rep resentations for visible-infrared person re-identification. IEEE Trans. Image Process. **29**, 579–590 (2019)
43. Ye, M., Shen, J., Lin, G., et al. Deep learning for person re-identification: a survey and outlook. arXiv: 2001.04193, (2020)
44. Li, D., Wei, X., Hong, X., et al.: Infrared- visible cross modal person re-identification with an X modality. In: 34th AAAI Conference on Artificial Intelligence, pp. 4610–4617 (2020)
45. Ye, M., Shen, J., Shao, L.: Visible- infrared person re identification via homogeneous augmented tri-modal learning. IEEE Trans. Inf. Forensics Secur. **16**, 728–739 (2020)
46. Jia, M., Zhai, Y., Lu, S., et al.: A similarity inference metric for RGB-infrared cross-modality person re-identification. arXiv: 2007.01504, (2020)
47. Liu, H., Cheng, J., Wang, W., et al.: Enhancing the discriminative feature learning for visible-thermal cross-modality person re-identification. Neurocomputing **398**, 11–19 (2020)
48. Ye, M., Lan, X.Y., Leng, Q.M., Shen, J.B.: Cross-modality person re-identification via modality-aware collaborative ensemble learning. IEEE Trans. Image Process. **29**, 9387–9399 (2020). https://doi.org/10.1109/TIP.2020.2998275

Xiongjun Wen (b. 1978) received the B.S degree in Mechanical Automation from Xingjiang University in 2000, and received the M.S. degree in Pattern recognition and intelligent system from Central South Universityin 2011. Now, she is a associate professor at Information and electromechanical engineering, Hunan International Economics University, China. The research interests include information technology and e-commerce.



Xin Feng (b.1982) received the B.S degree in in computer science and technology from Hunan Normal University in 2004, and received the M.S. degree in computer application from Central South University in 2011. Now, Now, he is a lecturer at Information and electromechanical engineering, Hunan International Economics University, China. Research interests include data science and information technology.

Ping Li (b.1979) received the B.S degree in information technology education from Hunan Normal University in 2007, and received the M.S. degree in computer application from Central South University in 2011. Now, Now, he is a associate professor at Information and electromechanical engineering, Hunan International Economics University, China. Research interests include network technology and information security.

Wenfang Chen (b. 1980) received the B.S degree in Computer Science and Technology from Hunan Normal University in 2003, and received the M.S. degree in Computer Science and Technology from Hunan University. Now, she is a researcher at College of Computer, Hunan University of International Economics, China. The research interests include image processing and information security.