Leaning compact and representative features for cross-modality person re-identification

Guangwei Gao[1,2] · Hao Shao[1] · Fei Wu[1] · Meng Yang[3] · Yi Yu[2]

Abstract

This paper pays close attention to the cross-modality visible-infrared person re-identification (VI Re-ID) task, which aims to match pedestrian samples between visible and infrared modes. In order to reduce the modality-discrepancy between samples from different cameras, most existing works usually use constraints based on Euclidean metric. Because of the Euclidean based distance metric strategy cannot effectively measure the internal angles between the embedded vectors, the existing solutions cannot learn the angularly discriminative feature embedding. Since the most important factor affecting the classification task based on embedding vector is whether there is an angularly discriminative feature space, in this paper, we present a new loss function called Enumerate Angular Triplet (EAT) loss. Also, motivated by the knowledge distillation, to narrow down the features between different modalities before feature embedding, we further present a novel Cross-Modality Knowledge Distillation (CMKD) loss. Benefit from the above two considerations, the embedded features are discriminative enough in a way to tackle modality-discrepancy problem. The experimental results on RegDB and SYSU-MM01 datasets have demonstrated that the proposed method is superior to the other most advanced methods in terms of impressive performance. Code is available at https://github.com/IVIPLab/LCCRF.

1 Introduction

Image retrieval is a research hotspot in computer vision community [20]. Among them, person re-identification refers to matching pedestrian images acquired from disjoint cameras [8, 9, 13]. In recent years, it has received substantial attention due to its significant practical value

---

Guangwei Gao and Hao Shao contributed equally to this work.

✉ Guangwei Gao

Extended author information available on the last page of the article

in video surveillance [14, 43]. Conventional person re-identification is only devoted to single-modality,i.e. all the person images are taken by visible cameras during daytime. Neverthe-less, the visible cameras cannot image clearly in the dark environment, which impedes the popularization and application of person re-identification [32]. To overcome this obstacle, in addition to the visible cameras, infrared cameras that are robust to illumination variants are also equipped in many surveillance scenarios. Therefore, in practice, we often need to match visible (RGB) and Infrared (IR) pedestrian images.

To narrow the modality gap between the infrared and visible images, recent works [3, 29] used Euclidean metric-based constraints [1, 16] to force the features from the same identity to be closer than those from the different identities. Although the above approaches have been successful, there still exists an inherent defect in the design of the loss function based on the Euclidean measure: the angle between embedded features cannot be effectively constrained by triple loss, which leads to the indistinguishable direction of features in common space. In the training stage, the included angles of the negative sample pairs may be smaller than that of the positive sample pairs, which makes the model impossible to divide the appropriate area for features to be embedded in the common space. Therefore, the angularly discriminative feature space that we expect is often not available.

More importantly, an angularly discriminative feature space is crucial to the final classifica-tion loss in the liner layer. It is calculating the dot product between feature vectors and weight vectors. When the weight vectors are assured, then the result is determined by the included angles of the feature vectors only. To solve the above problems, we design a loss function named Enumerate Angular Triplet (EAT) loss, which focuses on the included Angle between the embedded vectors generated by different modes and uses the cosine distance to measure the included Angle between the embedded vectors. Most previous methods first learned the unique features of the modal in the feature extraction stage, then mapped the features between different modalities into a common space and narrowed the distance between them. But they overlooked an important issue, the insurmountable gap between the different modalities still exists. Therefore, in the feature extraction stage, extracting the unique features of the modal while reducing the distance between them is authentically beneficial to the subsequent feature embedding stage. To this end, we also propose a new loss function, named Cross-Modality Knowledge Distillation (CMKD) loss, to narrow the distance between different modal features at the end of the unique feature extraction stage.

Our main contributions can be listed as three-fold:

- We devise an efficient Enumerate Angular Triplet (EAT) loss, which can better help to obtain an angularly separable common feature space via explicitly restraining the internal angles between different embedding features, contributing to the improvement of the per-formance.
- Motivated by the knowledge distillation, a novel Cross-Modality Knowledge Distillation (CMKD) loss is proposed to reduce the modality discrepancy in the modality-specific fea-ture extraction stage, contributing to the effectiveness of the cross-modality person Re-ID task.
- Our network achieves prominent results on both SYSU-MM01 and RegDB datasets with-out any other data augment skills. It achieves a Mean Average Precision (mAP) of 43.09% and 79.92% on SYSU-MM01 and RegDB datasets, respectively.

2 Related work

2.1 RGB Re-ID

The grateful appearances-based RGB Re-ID approaches [52] mainly focus on how to better learn the semantic representation of high-level features (such as attributes and depth features) or low-level ones (such as shape, color, and texture) that are more discriminative. Along with rapid development of convolutional neural networks, deep-based solutions have achieved promising performance recently. Person Re-ID approaches combined with supervised deep learning can usually be divided into two categories: One is the method based on representation learning, the other is the method based on measurement learning. In the representation-based learning approach, person Re-ID is usually considered as a visual classification problem and the similarity is calculated by using the embedding features projected to the common feature representation space. The researchers hope that these embedding features existing in common space can better describe person images after using human labels in the training stage. Recent works include part based [26, 51], GAN based [23, 34], and attention based [17, 43] ones, etc.

The metric learning based methods can make the model better learn the potential correlation between the data, which is beneficial to the model to learn more discriminant features. The metric learning steered methods aim at learning discriminative features by preform the feature distances comparison between distinguishing samples. Some popular metric learning steered methods [15, 30, 49, 50] proposed well-designed loss functions like triplet loss and contrastive loss to reduce the distances of features from the same identity and meanwhile magnify those from different identities. Angular constraints, also named cosine softmax loss [35], have been well studied in RGB Re-ID community. The cosine softmax loss is mainly designed by combining L2 normalization with complex classification loss based on softmax to achieve the purpose of enhancing the angle of embedded features and making it easier to distinguish. On the contrary, our proposed approach focuses on designing a more concise ranking loss that directly constrain the angle of the resulting embedded features.

2.2 RGB-IR Re-ID

For the demand of security in the dark, although infrared camera has been an important part of visual information acquisition, there are still few other studies on cross-modality RGB-IR Re-ID. Recently, to better study cross-modality RGB-IR Re-ID, the researchers proposed a large-scale dataset called SYSU-MM01 [36]. In such dataset, the query set images are from the infrared camera, while the gallery set images are from the visible camera, which is more realistic.

Most recent works on cross-modality RGB-IR Re-ID can be divided into two categories: single-stream-based and two-stream-based networks. By padding the multi-modality inputs into domain-specific nodes, Wu et al. [36] proposed a deep zero-padding strategy based single-stream model to match cross-modality features. Wang et al. [29] and Dai et al. [3] considered the adversarial training scheme, and designed a special pipeline to guide model learning and achieved remarkable performance. In the related studies based on the two-stream models, HCML [40] and BDTR [42] combined the representation learning and measurement learning simultaneously to constrain the model. On the basis of the

two-stream network, MSR [7] integrated the view classification [6] scheme and proposed a new network to learn the modality speciality and modality sharing features.

2.3 Cross-Modality Feature Learning

Cross-modal retrieval matches the modalities of input queries with the output results from different modalities. The main goal of cross-modal retrieval is to mitigate the "modality gap" caused by the inconsistent feature representations between different modalities [10, 25]. The multilayer deep neural networks (DNNs) based methods have been widely designed to project the heterogeneous features into a consistent semantic space where the feature metric learning can be performed.

For example, Xu et al. [38] proposed cross-modal attention with semantic consistency to perform cross-modality feature embedding for image-text matching task. For effective audio-visual association, Zhang et al. [47] designed a self-supervised curriculum learning method in terms of the teacher-student learning framework. The contrastive learning scheme is explored to distill and capture the cross-modal correspondence. Lu et al. [19] presented a cross-modality shared-specific feature transfer algorithm, which can perform discriminative and complementary feature learning. Ye et al. [45] devised a Homogeneous Augmented Tri-Modal learning method, which performs tri-modal feature learning to reduce cross-modality variations. Wei et al. [33] proposed an adversarial learning-based flexible body partition model to alleviate the cross-modality gap and promote the feature representation capability.

3 Methods

In this part, we will detail our presented method for cross-modality person re-identification, as elaborated in Figure 1. The following description mainly includes three parts: (1) the backbone network, (2) the Enumerate Angular Triplet (EAT) loss, and (3) the Cross-modality Knowledge Distillation (CMKD) loss.

3.1 Network backbone

The most commonly used method in the field of visible-thermal cross-modality person re-identification problem is a two-stream network, which was first introduced in [42]. The network is composed of feature extraction and feature embedding. The purpose of the feature extraction is to learn modality-specific information of visible and infrared modality, while the target of the feature embedding is to learn the modality-shared common features between the above two modalities.

In the existing work, the first part (feature extraction) is often implemented directly through some designed convolution neural networks, such as ResNet50, etc, while the second part (feature embedding) is usually implemented by some commonly shared fully connected layers. Feature extraction is mainly composed of two branches that do not share the parameters. But if each branch contains a whole CNN architecture, the number of network parameters can be multiplied. Feature embedding is mainly composed of several full-linked layers with shared parameters. But it can only deal with the 1D-shaped feature vectors and ignore critical personal space structure information about a person.
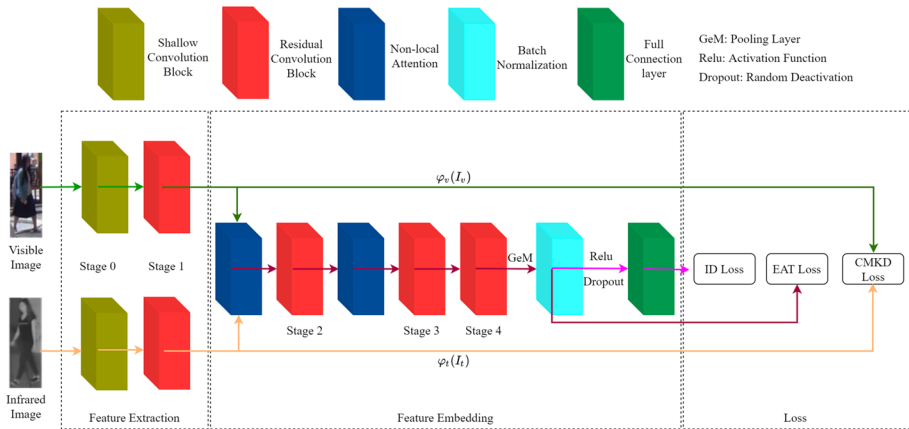
Fig. 1 The pipeline of our proposed method for cross-modality person Re-ID. The network mainly consists of two stages. In the first stage, the feature extraction part is composed of two independent branches that do not share parameters, which is used to learn the unique features of the two modes. In the second stage, the feature mapping part is composed of a common network with shared parameters, which is used to map the learned features to a public space. The experiment has three constraints: (1). The proposed Enumerate Angular Triplet (EAT) loss; (2). The proposed Knowledge Distillation (KD) loss; (3). Identity (ID) Loss

To handle the above two issues, we take full advantage of the previous experience and divide the convectional CNN model into two parts. ResNet50 is mainly composed of the shallow convolution block stage0 and those res-convolution blocks stage1, stage2, stage3, and stage4. We use stage0 and stage1 as the feature extraction part and stage2, stage3, and stage4 as the next feature embedding part.

For the sake of description, we use φ_v and φ_t to represent visible lightweight feature extraction function and infrared feature extraction function, respectively. The above two networks are used to attain modality-specific information. The following feature embedding network is represented by φ_{vt}, which can project the learned features into a common feature representation space. When acquired a visible image I_v and an infrared image I_t, the final feature learned in the common space is described as

$$V = \varphi_{vt}(\varphi_v(I_v)). \tag{1}$$

$$T = \varphi_{vt}(\varphi_t(I_t)). \tag{2}$$

The attention scheme has been proven that they can play a vital role in cross-modality Re-ID tasks. We leverage the simple yet effective non-local attention block in [28] to attain the more meaningful descriptions of all positional features, which is represented by

$$z_i = W_z * \varphi(x_i) + x_i, \tag{3}$$

where $\varphi(.)$ denotes a non-local operation, W_z is the desired weight matrix to be learned and $+x_i$ formulates a residual learning scheme.

In the person re-identification tasks, neither the most frequently used average pool operation nor the maximum pool operation can capture the domain-specific distinguishing features. Therefore, we utilize a generalized-men (GeM) [21, 24] pooling layer. The feature mapping of 3D parts is transformed into the feature vector of 1D parts. Given an

intermediate 3D feature description $X \in R^{C \times W \times H}$, the popular GeM can be represented as

$$X = \left(\frac{1}{|X|} \sum_{x_i \in X} x^p_i \right)^{\frac{1}{p}}, \tag{4}$$

where $X \in R^{C \times 1 \times 1}$ denotes the desired pooled results, |.| is the element amount, p denotes the pooling hyper-parameter, that can be pre-set or learned by the back-propagating. When $p \to \infty$, GeM approximates max-pooling. While when $p \to 1$, GeM approximates average-pooling.

3.2 Cross-modality knowledge distillation loss

First, we consider narrowing the distance between different modal features in the feature extraction stage to make the following feature embedding more reasonable and effective. At present, most previous methods first learn the unique features of the modal in the feature extraction stage and then map the features between different modalities into a common space after the feature embedding stage. But they have overlooked an important issue, the insurmountable gap between modalities still exists. Therefore, in the feature extraction stage, extracting the unique features of the modality while reducing the distance between the modalities is beneficial to the subsequent work of the feature embedding stage. To this end, as shown in Figure 2, in the feature extraction stage, we propose a novel loss function, called Cross-Modality Knowledge Distillation (CMKD) loss as follows:
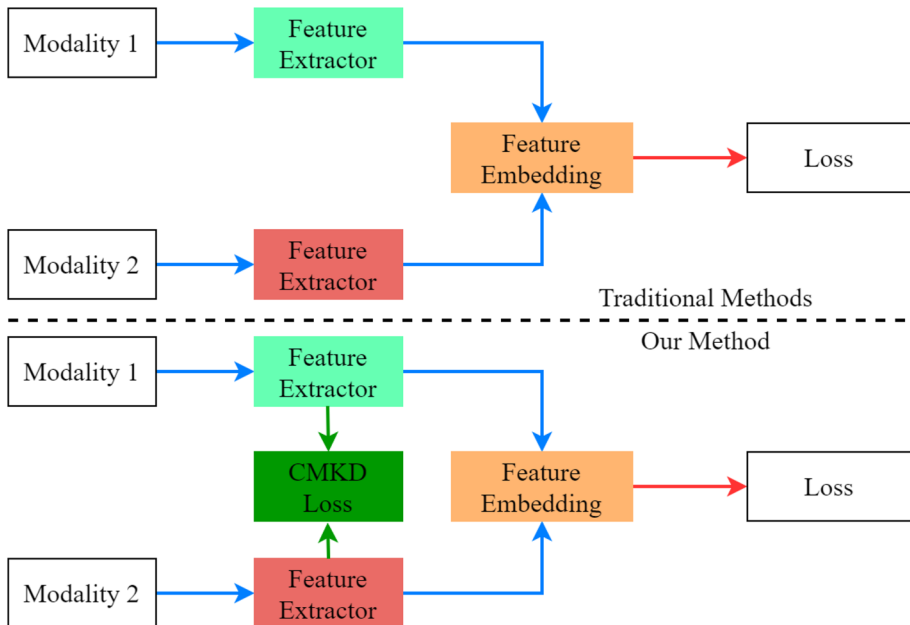

Fig. 2 Comparisons between the conventional networks and our designed network. Through our experiments, we found that CMKD loss is beneficial to obtain more discriminant features in the feature embedding space

$$L_{CMKD} = \frac{1}{N} \sum_{I_a, I_p} ||V^{rgb}_a - V^{ir}_p||_2^2 + ||V^{ir}_a - V^{rgb}_p||_2^2, \tag{5}$$

where V^{rgb}_a and V^{rgb}_p is the embedding features of anchor sample and positive sample in RGB modality, V^{ir}_a and V^{ir}_p is the embedding features of anchor sample and positive sample in infrared modality, and $||.||_2$ denotes the Euclidean distance.

3.3 Enumerate angular triplet loss

The triple loss function is designed to make the distance between the anchor image I_a and the positive image I_p closer than that between the anchor sample I_a and the negative sample I_n by a constraint ζ.

As the triplet loss is difficult to separate the orientation of the embedding features, we put forward a novel bi-directional Enumerate Angular Triplet (EAT) loss function. Our proposed loss function uses not Euclidean distance but cosine distance to represent the similarities between embedded features.

Taking those above factors into account, we first design an efficient loss function named angular triple (AT) loss as follows:

$$L_{\cos} = \frac{1}{N} \sum_{I_a, I_p, I_n} [\cos(V_a, V_n) - \cos(V_a, V_p) + \zeta]_+, \tag{6}$$

where N is the number of the identity class, V_a is the embedding features of the anchor sample I_a, V_n is the embedding features of the negative sample I_n, and V_p is the embedding features of the positive sample I_p.

However, the proposed function suffers from two drawbacks. First, it is our target to promote V_a and V_n easier to be distinguished in the embedded feature representation space. Therefore, there needs to be a clamping function to achieve the above purpose. Second, the overall clamping function is excluded with an appropriate selection of margin ζ. When the margin is not less than 1, the loss function is usually non-negative. Therefore, we set the margin to be 1 to maintain the parameters uncomplicated.

Given the above considerations, we then reformulate the AT loss function as follows:

$$L_{AT} = \frac{1}{N} \sum_{I_a, I_p, I_n} ([\cos(V_a, V_n)]_+ - [\cos(V_a, V_p)]_+ + 1). \tag{7}$$

Regarding the cross-modality enumeration loss function, we first consider the inter-class cross-modality constraints L_{crgb} and L_{cir}, which are similar to the direct triplet loss:

$$L_{crgb} = \frac{1}{N} \sum_{I_a, I_p, I_n} \{([\cos(V^{rgb}_a, V^{ir}_n)]_+ \\ - [\cos(V^{rgb}_a, V^{ir}_p)]_+ + 1\}, \tag{8}$$

$$L_{cir} = \frac{1}{N} \sum_{I_a, I_p, I_n} \{([\cos(V^{ir}_a, V^{rgb}_n)]_+ \\ - [\cos(V^{ir}_a, V^{rgb}_p)]_+ + 1\}. \tag{9}$$

Based on the above formulations, the inter-class same-modality constraints L_{srgb} and L_{sir} are designed to mitigate the modality gap at the image patch level:

$$L_{srgb} = \frac{1}{N} \sum_{I_a, I_p, I_n} \{([\cos(V^{rgb}{}_a, V^{rgb}{}_n)]_+ \\ - [\cos(V^{rgb}{}_a, V^{ir}{}_p)]_+ + 1\}, \tag{10}$$

$$L_{sir} = \frac{1}{N} \sum_{I_a, I_p, I_n} \{([\cos(V^{ir}{}_a, V^{ir}{}_n)]_+ \\ - [\cos(V^{ir}{}_a, V^{rgb}{}_p)]_+ + 1\}. \tag{11}$$

In summary, we can formulate the bi-directional extension of AT loss as follows:

$$L_{ATrgb} = L_{crgb} + L_{srgb}, \tag{12}$$

$$L_{ATir} = L_{cir} + L_{sir}, \tag{13}$$

$$L_{EAT} = L_{ATrgb} + L_{ATir}. \tag{14}$$

Moreover, since the exponential function $y = e^x$ grows exponentially at $x > 0$, this characteristic is beneficial to the rapid convergence of the model. Thus, we give the bi-directional enumerate AT loss as

$$L_{EATrgb} = \frac{1}{N} \sum_{I_a, I_p, I_n} e^{([\cos(V^{rgb}{}_a, V^{ir}{}_n)]_+ - [\cos(V^{rgb}{}_a, V^{ir}{}_p)]_+ + 1)} \\ + \frac{1}{N} \sum_{I_a, I_p, I_n} e^{([\cos(V^{rgb}{}_a, V^{rgb}{}_n)]_+ - [\cos(V^{rgb}{}_a, V^{ir}{}_p)]_+ + 1)}, \tag{15}$$

$$L_{EATir} = \frac{1}{N} \sum_{I_a, I_p, I_n} e^{([\cos(V^{ir}{}_a, V^{rgb}{}_n)]_+ - [\cos(V^{ir}{}_a, V^{rgb}{}_p)]_+ + 1)} \\ + \frac{1}{N} \sum_{I_a, I_p, I_n} e^{([\cos(V^{ir}{}_a, V^{ir}{}_n)]_+ - [\cos(V^{ir}{}_a, V^{rgb}{}_p)]_+ + 1)}, \tag{16}$$

$$L_{EAT} = L_{EATrgb} + L_{EATir}. \tag{17}$$

However, in our experiments, we find that loss function (17) is difficult to be converged. In our designed enumeration loss, to ensure that each element of the generated feature description is as evenly distributed as possible, thus making the descriptor more informative and discriminative, a compactness term C is introduced. It is computed by comparing the differences between each element in $f(.)$ and the mean value:

$$C = \sum_{i}^{N} \sum_{r}^{R} e^{\cos(f_r(V^{rgb}{}_a), \bar{f}(V^{rgb}{}_a))} +$$
$$\sum_{i}^{N} \sum_{r}^{R} e^{\cos(f_r(V^{ir}{}_a), \bar{f}(V^{ir}{}_a))}. \tag{18}$$

Here, $f_r(.)$ and $\bar{f}(.)$ denotes the r-th element and the mean index of all elements in $f(.)$, respectively. R denotes the length of the acquired local deep descriptor $f(.)$. The introduction of C can also avoid overfitting in network training. In our experiment, if the compactness term in the loss is not enumerated, it is difficult for the network to converge.

In summary, our enumerate angular triplet loss function is defined by integrating them as follows:

$$L_{EAT} = L_{EAT} + C. \tag{19}$$

Figure 3 plots the two-dimensional visualization of the features in the common feature space with respective methods. We can observe that by using our proposed EAT loss which considers both intra-modal and inter-modal distance, the embedding features from different classes in the common feature space are drastically separable, which contributes to the following classification performance. On the contrary, the expAT loss only considers the distance between images from different modalities and ignores the intra-modal distance that can provide discriminative and complementary information.

3.4 Overall Loss

In addition, to achieve better classification results, similar to some of the advanced methods, we also take into account identity loss, which integrates identity-specific information by treating each person as one class. The formula for identity loss is given as follows:
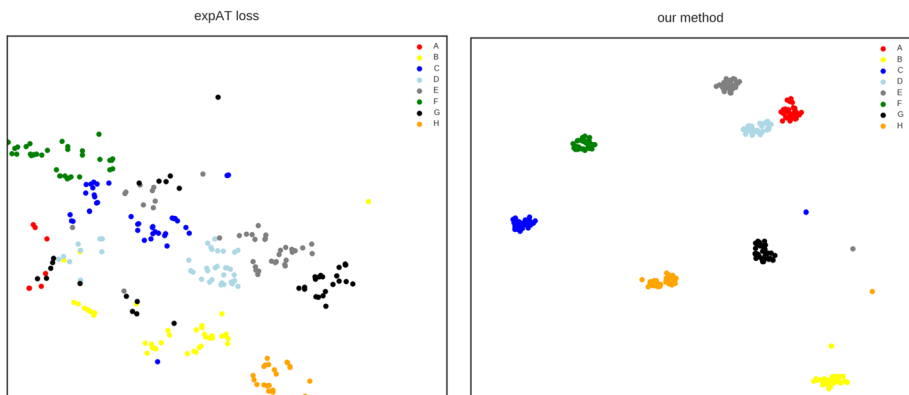

Fig. 3 Two-dimensional visualization of query subset features in the SYSU-MM01 dataset in the common feature space. Compared with expAT loss [46], the embedded vectors of different classes in our proposed method are easier to be separated in the feature space

$$L_{ID} = \sum_{i=1}^{N} -q_i \log(p_i), \tag{20}$$

where p_i is the prediction label of the ith class, q_i is the true label of the ith class, N is the number of all classes of the training samples.

So far, the final loss of our method is calculated as follows:

$$L_{ALL} = L_{EAT} + L_{CMKD} + L_{ID}. \tag{21}$$

4 Experiments

In this section, we conducted evaluations on RegDB [22] (providing infrared images by thermal cameras) and SYSU-MM01 [36] (providing infrared images by near-infrared cameras) datasets to verify the efficiency and effectiveness of our proposed method. The example images are listed in Figure 4.

4.1 Datasets

1) SYSU-MM01 set: The SYSU-MM01 dataset is a dataset consisting of 491 identities providing RGB and IR images from six cameras. The training set contains 395 people, including visible images 22,258 pieces and infrared images 11909 pieces. The other 96 people are included in the testing set, including 3803 infrared samples for query and 301 randomly selected visible samples as the gallery set.

2) RegDB set: The RegDB dataset is one of the most commonly used datasets in the RGB-thermal person re-identification field. It consists of 8,240 photos from 412 identities, each with 10 visible images and 10 infrared images.

4.2 Evaluation metrics

Based on previous work, Cumulative Matching Characteristics (CMC) and Mean Average Precision (mAP) are used as the evaluation indexes. To evaluate the accuracy of the results, we repeated the final evaluation 10 times, randomly segmented each time with different gallery sets, and calculated statistically stable performance indicators as the final results.

4.3 Implementation details

Experiments are carried out on two datasets to verify the proposed method, and the algorithm is implemented using Pytorch 1.2. The batch size is set to 8 and the proposed method is optimized using the ADAM optimizer [12] with a warm-up [5] strategy, with an initial learning rate set as 3×10^{-4} and the decay of 0.1 at 10,000 and 20,000 steps. The training process is iterated by 30,000 steps. For the anchor of both modalities, the margin of the triplet loss is empirically set as 0.3. We also use the Label Smoothing [27] and the Random Erasing [53] strategies to alleviate the overfitting problem in the training stage, and the corresponding parameters are set as 0.1 and 0.5 respectively. In the test, Euclidean distance is used to calculate the feature distance.

Fig. 4 Examples of cross-modality images. The images in the first part are from the RegDB set, where the first row shows the images captured by a thermal imaging camera and the second row shows the images captured by a visible light camera. The images in the second part are from the SYSU-MM01 set, where the first row shows the images captured by the infrared camera and the second row shows the image captured by the visible light camera. Two images in each column from the same identity

Batch Sampling Strategy: To better perform cross-modality constraints, we use a special sampling strategy. First, we use a visible sample and an infrared sample of the same identity as an image pair. For the batch size N, we randomly select N anchor sample pairs from the entire training set. Then, a negative RGB sample and a positive RGB sample are randomly selected from the training dataset where the RGB images of anchor points are eliminated, to form a tuple with the anchor. So there's a total of $N \times 2 \times 3$ samples in each mini-batch. In our experiment, anchor samples of the same identity from different modalities of the same tuple are explored to calculate the identity loss, and all images in the same tuple are explored to calculate the ranking loss. We use the random sampling strategy to traverse all the training samples in the training process.

4.4 Ablation Study

In this part, we evaluate and analyze the effectiveness of the proposed method through qualitative and quantitative tests on the popular SYSU-MM01 set. In this experiment, the results are reported with all search single-shot settings. Specifically, "EAT Loss" represents the Enumerate Angular Triplet (EAT) loss, "Non-Local Attention" represents the non-local attention block, and "CMKD Loss" represents the Cross-modality Knowledge Distillation (CMKD) loss.

We can make the following observations through the results shown in Table 1. (1) By using the network with EAT loss, we can achieve better performance than the two-stream network in [7, 32, 46]. This experiment demonstrates that an angularly informative and discriminative feature space is explicitly beneficial for cross-modality Re-ID. (2) By using CMKD loss, performance is improved by narrowing the distance between different modal features in the feature extraction stage. (3) When aggregating two losses with non-local attention, the performance is further improved, demonstrating that these losses and attention are mutually beneficial to each other.

4.5 Peer comparisons

In this section, we list and compare some state-of-the-art approaches, including eBDTR [41], HSME [11], D^{2}RL [29], MAC [39], MSR [7], AlignGAN [31], EDFL [18], and HPILN [48], etc. Most of the methods we compared were published in the last two years. The results performance on the popular SYSU-MM01 and REGDB sets are provided in Table 2 and Table 3, respectively.

The experimental results on SYSU-MM01 datasets demonstrate that our proposed method achieves the best performance in all query settings, achieving 43.23%/43.09% rank-1/mAP for the All Search setting and 50.07%/58.88% rank-1/mAP for the Indoor Search setting. The experimental results on RegDB dataset also demonstrate that our method achieves the best performance in both query settings, usually by a large margin, achieving 79.27%/77.69% rank-1/mAP for the RGB to thermal query setting and 80.97%/79.92% rank-1/mAP for the thermal to RGB query setting.

Also, no additional parameters are introduced into the proposed method during testing, which indicates that this method is easier to use in practical application scenarios. We also provide some retrieval results in Figures 5 and 6. A large number of above experiments

Table 1 Ablation studies (%) of the presented method on SYSU-MM01 set

EAT Loss	Non-Local Attention	CMKD Loss	SYSU-MM01			
			Rank1	Rank10	Rank20	mAP
✗	✗	✗	21.14	50.25	62.77	18.52
✓	✗	✗	40.17	78.73	87.64	40.11
✓	✓	✗	42.16	78.76	87.52	41.88
✓	✗	✓	41.80	80.07	89.26	41.29
✓	✓	✓	43.23	82.78	90.91	43.09

We study the effects of three different component combinations including EAT loss, non-local attention and CMKD loss

Table 2 Comparison performance (%) on the SYSU-MM01 set. "-" indicates not available or not provided.

Methods	All Search				Indoor Search			
	Rank1	Rank10	Rank20	mAP	Rank1	Rank10	Rank20	mAP
Zero-pad [36]	14.80	54.12	71.33	15.95	20.58	68.38	85.79	26.92
cmGAN [3]	26.97	67.51	80.56	27.80	31.63	77.23	89.18	42.19
HCML [40]	14.32	53.16	69.17	16.16	24.52	73.25	86.73	30.08
HSME [11]	20.68	62.74	77.95	23.12	–	–	–	–
D^{2}RL [29]	28.90	70.60	82.40	29.20	–	–	–	–
MAC [39]	33.26	79.04	90.09	36.22	36.43	62.36	71.63	37.03
AlignGAN [31]	42.40	85.00	93.70	40.70	45.90	87.60	94.40	54.30
HPLIN [48]	41.36	84.78	94.51	42.95	45.77	91.82	98.46	56.52
Hi-CMD [2]	34.94	77.58	–	35.94	–	–	–	–
EDFL [18]	36.94	85.42	93.22	40.77	–	–	–	–
CDP [4]	38.00	82.30	91.70	38.40	–	–	–	–
expAT [46]	38.57	76.64	86.39	38.61	–	–	–	–
eBDTR [44]	27.82	67.34	81.34	28.42	32.46	77.42	89.62	42.46
MSR [7]	37.35	83.40	93.34	38.11	39.64	89.29	97.66	50.88
JSIA [32]	38.10	80.70	89.90	36.90	43.80	86.20	94.20	52.90
Ours	43.23	82.78	90.91	43.09	50.07	90.63	96.99	58.88

Table 3 Peer comparisons (%) on the RegDB set under "Thermal to RGB" and "RGB to Thermal" experimental settings

Methods	Thermal to RGB		RGB to Thermal	
	Rank1	mAP	Rank1	mAP
Zero-Pad [36]	16.63	17.82	17.75	31.83
HCML [40]	21.70	22.24	24.44	20.08
D^{2}RL [29]	–	–	43.40	44.10
MSR [7]	–	–	48.43	48.67
AlignGAN [32]	56.30	53.40	–	–
EDFL [18]	51.89	52.13	52.58	52.98
AGW [41]	–	–	70.05	66.37
CMSP [37]	–	–	65.07	64.50
expAT [46]	67.45	66.51	66.48	67.31
Hi-CMD [2]	–	–	70.93	66.04
cm-SSFT [19]	71.00	71.70	72.30	72.90
HAT [45]	70.02	66.30	71.83	67.56
FBP-AL [33]	70.05	66.61	73.98	68.24
Ours	80.97	79.92	79.27	77.69

Fig. 5 Top-5 retrieval results on popular SYSU-MM01 set under "Infrared to RGB" setting. For each query on the left, top-*k* candidates are listed in ascending order due to their similarities. The false and true retrievals are given in the red and green boxes, respectively

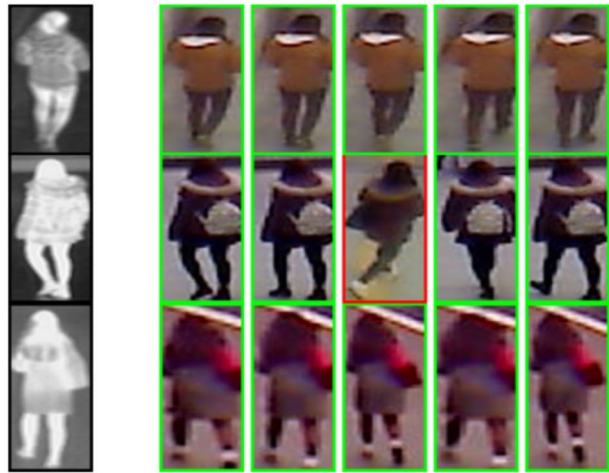(a) Top-5 retrieval results from our method

(b) Top-5 retrieval results from expAT method [33]

have shown that cross-modal shared feature representation for cross-modal person re-recognition tasks can be better learned by using our proposed EAT loss and CMKD loss.

5 Conclusions

The purpose of this paper is to improve discriminative feature learning by a simple method. On one hand, motivated by the knowledge distillation, a new Cross-Modality Knowledge Distillation (CMKD) loss is explicitly presented to reduce the modality discrepancy in the modality-specific feature extraction stage. On the other hand, in order to help the deep network learn angularly representative embedded features from different modalities, we put forward the Enumerate Angular Triplet (EAT) loss. The EAT loss can constrain the included angle between the embedded vectors, which is helpful for angular segmentation

Fig. 6 Top-5 retrieval results on popular RegDB dataset under "Thermal to RGB" setting. For each query on the left, top-*k* candidates are listed in ascending order due to their similarities. The false and true retrievals are given in the red and green boxes, respectively

(a) Top-5 retrieval results from our method
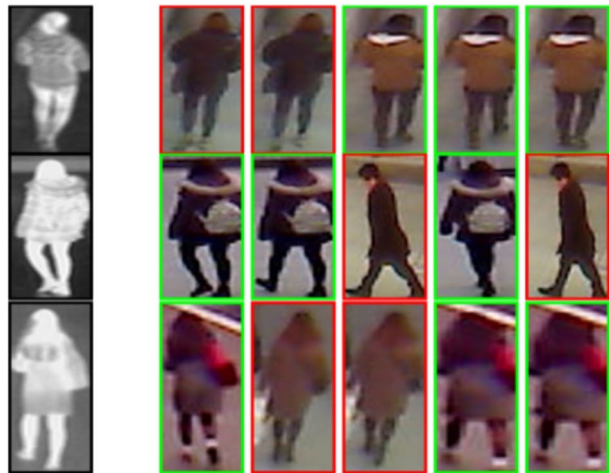
(b) Top-5 retrieval results from expAT method [33]

of the feature space. Experimental results on two cross-modality Re-ID datasets have shown that the proposed method is effective compared with the most advanced methods.

Declarations

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

References

1. Chen, W., Chen, X., Zhang, J., Huang, K.: Beyond triplet loss: a deep quadruplet network for person re-identification. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 403–412 (2017)

2. Choi, S., Lee, S., Kim, Y., Kim, T., Kim, C.: Hi-cmd: hierarchical cross-modality disentanglement for visible-infrared person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10257–10266 (2020)

3. Dai, P., Ji, R., Wang, H., Wu, Q., Huang, Y.: Cross-modality person re-identification with generative adversarial training. In: International Joint Conference on Artificial Intelligence, pp. 677–683 (2018)

4. Fan, X., Luo, H., Zhang, C., Jiang, W.: Cross-spectrum dual-subspace pairing for rgb-infrared cross-modality person re-identification. arXiv:2003.00213 (2020)

5. Fan, X., Jiang, W., Luo, H., Fei, M.: Spherereid: Deep hypersphere manifold embedding for person re-identification. Journal of Visual Communication and Image Representation **60**, 51–58 (2019)

6. Feng, Z., Lai, J., Xie, X.: Learning view-specific deep networks for person re-identification. IEEE Transactions on Image Processing **27**(7), 3472–3483 (2018)

7. Feng, Z., Lai, J., Xie, X.: Learning modality-specific representations for visible-infrared person re-identification. IEEE Transactions on Image Processing **29**, 579–590 (2020)

8. Gao, G., Yu, Y., Yang, J., Qi, G.J., Yang, M.: Hierarchical deep cnn feature set-based representation learning for robust cross-resolution face recognition. IEEE Transactions on Circuits and Systems for Video Technology (2020)

9. Gao, G., Yang, J., Jing, X.Y., Shen, F., Yang, W., Yue, D.: Learning robust and discriminative low-rank representations for face recognition with occlusion. Pattern Recognition **66**, 129–143 (2017)

10. Gao, G., Yu, Y., Xie, J., Yang, J., Yang, M., Zhang, J.: Constructing multilayer locality-constrained matrix regression framework for noise robust face super-resolution. Pattern Recognition **110**, 107539 (2021)

11. Hao, Y., Wang, N., Li, J., Gao, X.: Hsme: hypersphere manifold embedding for visible thermal person re-identification. In: Proceedings of the AAAI conference on Artificial Intelligence, pp. 8385–8392 (2019)

12. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv:1412.6980 (2014)

13. Leng, Q., Ye, M., Tian, Q.: A survey of open-world person re-identification. IEEE Transactions on Circuits and Systems for Video Technology **30**(4), 1092–1108 (2020)

14. Li, R., Zhang, B., Kang, D.J., Teng, Z.: Deep attention network for person re-identification with multi-loss. Computers & Electrical Engineering **79**, 106455 (2019)

15. Liu, H., Shi, W., Huang, W., Guan, Q.: A discriminatively learned feature embedding based on multi-loss fusion for person search. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1668–1672. IEEE (2018)

16. Liu, J., Zha, Z.J., Tian, Q., Liu, D., Yao, T., Ling, Q., Mei, T.: Multi-scale triplet cnn for person re-identification. In: Proceedings of the ACM international conference on Multimedia, pp. 192–196 (2016)

17. Liu, H., Feng, J., Qi, M., Jiang, J., Yan, S.: End-to-end comparative attention networks for person re-identification. IEEE Transactions on Image Processing **26**(7), 3492–3506 (2017)

18. Liu, H., Cheng, J., Wang, W., Su, Y., Bai, H.: Enhancing the discriminative feature learning for visible-thermal cross-modality person re-identification. Neurocomputing **398**, 11–19 (2020)

19. Lu, Y., Wu, Y., Liu, B., Zhang, T., Li, B., Chu, Q., Yu, N.: Cross-modality person re-identification with shared-specific feature transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13379–13389 (2020)

20. Lu, H., Zhang, M., Xu, X., Li, Y., Shen, H.T.: Deep fuzzy hashing network for efficient image retrieval. IEEE Transactions on Fuzzy Systems **29**(1), 166–176 (2020)

21. Luo, H., Jiang, W., Gu, Y., Liu, F., Liao, X., Lai, S., Gu, J.: A strong baseline and batch normalization neck for deep person re-identification. IEEE Transactions on Multimedia **22**(10), 2597–2609 (2019)

22. Nguyen, D.T., Hong, H.G., Kim, K.W., Park, K.R.: Person recognition system based on a combination of body images from visible light and thermal cameras. Sensors **17**(3), 605 (2017)

23. Qian, X., Fu, Y., Xiang, T., Wang, W., Qiu, J., Wu, Y., Jiang, Y.G., Xue, X.: Pose-normalized image generation for person re-identification. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 650–667 (2018)

24. Radenović, F., Tolias, G., Chum, O.: Fine-tuning cnn image retrieval with no human annotation. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(7), 1655–1668 (2019)

25. Serikawa, S., Lu, H.: Underwater image dehazing using joint trilateral filter. Computers & Electrical Engineering **40**(1), 41–50 (2014)

26. Sun, Y., Zheng, L., Yang, Y., Tian, Q., Wang, S.: Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 480–496 (2018)

27. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 2818–2826 (2016)

28. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 7794–7803 (2018)

29. Wang, Z., Wang, Z., Zheng, Y., Chuang, Y.Y., Satoh, S.: Learning to reduce dual-level discrepancy for infrared-visible person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 618–626 (2019)

30. Wang, G., Yuan, Y., Chen, X., Li, J., Zhou, X.: Learning discriminative features with multiple granularities for person re-identification. In: Proceedings of the ACM international conference on Multimedia, pp. 274–282 (2018)

31. Wang, G., Zhang, T., Cheng, J., Liu, S., Yang, Y., Hou, Z.: Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3623–3632 (2019)

32. Wang, G.A., Zhang, T., Yang, Y., Cheng, J., Chang, J., Liang, X., Hou, Z.G.: Cross-modality paired-images generation for rgb-infrared person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 12144–12151 (2020)

33. Wei, Z., Yang, X., Wang, N., Gao, X.: Flexible body partition-based adversarial learning for visible infrared person re-identification. IEEE Transactions on Neural Networks and Learning Systems (2021)

34. Wei, L., Zhang, S., Gao, W., Tian, Q.: Person transfer gan to bridge domain gap for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 79–88 (2018)

35. Wojke, N., Bewley, A.: Deep cosine metric learning for person re-identification. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 748–756. IEEE (2018)

36. Wu, A., Zheng, W.S., Yu, H.X., Gong, S., Lai, J.: Rgb-infrared cross-modality person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5380–5389 (2017)

37. Wu, A., Zheng, W.S., Gong, S., Lai, J.: Rgb-ir person re-identification by cross-modality similarity preservation. International Journal of Computer Vision **128**(6), 1765–1785 (2020)

38. Xu, X., Wang, T., Yang, Y., Zuo, L., Shen, F., Shen, H.T.: Cross-modal attention with semantic consistence for image-text matching. IEEE Transactions on Neural Networks and Learning Systems **31**(12), 5412–5425 (2020)

39. Ye, M., Lan, X., Leng, Q.: Modality-aware collaborative learning for visible thermal person re-identification. In: Proceedings of the ACM International Conference on Multimedia, pp. 347–355 (2019)

40. Ye, M., Lan, X., Li, J., Yuen, P.: Hierarchical discriminative learning for visible thermal person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 7501–7508 (2018)

41. Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., Hoi, S.C.: Deep learning for person re-identification: A survey and outlook. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)

42. Ye, M., Wang, Z., Lan, X., Yuen, P.C.: Visible thermal person re-identification via dual-constrained top-ranking. In: International Joint Conference on Artificial Intelligence, vol. 1, p. 2 (2018)

43. Ye, M., Lan, X., Leng, Q., Shen, J.: Cross-modality person re-identification via modality-aware collaborative ensemble learning. IEEE Transactions on Image Processing **29**, 9387–9399 (2020)

44. Ye, M., Lan, X., Wang, Z., Yuen, P.C.: Bi-directional center-constrained top-ranking for visible thermal person re-identification. IEEE Transactions on Information Forensics and Security **15**, 407–419 (2020)

45. Ye, M., Shen, J., Shao, L.: Visible-infrared person re-identification via homogeneous augmented tri-modal learning. IEEE Transactions on Information Forensics and Security **16**, 728–739 (2020)

46. Ye, H., Liu, H., Meng, F., Li, X.: Bi-directional exponential angular triplet loss for rgb-infrared person re-identification. IEEE Transactions on Image Processing **30**, 1583–1595 (2021)

47. Zhang, J., Xu, X., Shen, F., Lu, H., Liu, X., Shen, H.T.: Enhancing audio-visual association with self-supervised curriculum learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 3351–3359 (2021)

48. Zhao, Y.B., Lin, J.W., Xuan, Q., Xi, X.: Hpiln: a feature learning framework for cross-modality person re-identification. IET Image Processing **13**(14), 2897–2904 (2019)

49. Zhao, C., Lv, X., Zhang, Z., Zuo, W., Wu, J., Miao, D.: Deep fusion feature representation learning with hard mining center-triplet loss for person re-identification. IEEE Transactions on Multimedia **22**(12), 3180–3195 (2020)

50. Zhao, C., Wang, X., Zuo, W., Shen, F., Shao, L., Miao, D.: Similarity learning with joint transfer constraints for person re-identification. Pattern Recognition **97**, 107014 (2020)
51. Zheng, W.S., Li, X., Xiang, T., Liao, S., Lai, J., Gong, S.: Partial person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4678–4686 (2015)
52. Zheng, L., Yang, Y., Hauptmann, A.G.: Person re-identification: Past, present and future. arXiv:1610.02984 (2016)
53. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 13001–13008 (2020)

Authors and Affiliations

Guangwei Gao[1,2] · **Hao Shao[1]** · **Fei Wu[1]** · **Meng Yang[3]** · **Yi Yu[2]**

Hao Shao
sh_0307@163.com

Fei Wu
wufei_8888@126.com

Yi Yu
yiyu@nii.ac.jp

[1] Institute of Advanced Technology, Nanjing University of Posts and Telecommunications, Nanjing, China

[2] Digital Content and Media Sciences Research Division, National Institute of Informatics, Tokyo, Japan

[3] Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, Sun Yat-sen University, Guangzhou, China