HSME: Hypersphere Manifold Embedding for Visible Thermal Person Re-Identification

Yi Hao,[†] Nannan Wang,[‡*] Jie Li,[†] Xinbo Gao[†]
[†]State Key Laboratory of Integrated Services Networks,
School of Electronic Engineering, Xidian University, Xi'an 710071, China
[‡]State Key Laboratory of Integrated Services Networks,
School of Telecommunications Engineering, Xidian University, Xi'an 710071, China

Abstract

Person Re-identification(re-ID) has great potential to contribute to video surveillance that automatically searches and identifies people across different cameras. Heterogeneous person re-identification between thermal(infrared) and visible images is essentially a cross-modality problem and important for night-time surveillance application. Current methods usually train a model by combining classification and metric learning algorithms to obtain discriminative and robust feature representations. However, the combined loss function ignored the correlation between classification subspace and feature embedding subspace. In this paper, we use Sphere Softmax to learn a hypersphere manifold embedding and constrain the intra-modality variations and cross-modality variations on this hypersphere. We propose an end-to-end dual-stream hypersphere manifold embedding network(HSMEnet) with both classification and identification constraint. Meanwhile, we design a two-stage training scheme to acquire decorrelated features, we refer the HSME with decorrelation as D-HSME. We conduct experiments on two cross-modality person re-identification datasets. Experimental results demonstrate that our method outperforms the state-of-the-art methods on two datasets. On RegDB dataset, rank-1 accuracy is improved from 33.47% to 50.85%, and mAP is improved from 31.83% to 47.00%.

1.Introduction

Person re-identification (ReID) aims at identifying a person from non-overlapping camera views, which has important value in video surveillance area. For example, given a query/probe pedestrian image, we need to retrieve all images of the same person ID in gallery images. A large number of algorithms for Re-ID problem have been proposed, such as (Zheng, Yang, and Hauptmann 2016) (Zheng, Zheng, and Yang 2017). Recent researches mainly focus on visible pedestrian images (Zheng et al. 2017)(Zheng et al. 2015)(Wang et al. 2018b)(Wang et al. 2016), i.e., both query images and gallery images are captured by visible camera. However, in night time or dark environment, visible images become uninformative. In such case, imaging devices that do not rely on visible light should be applied, which makes heterogeneous person re-identification significant for public surveillance applications.

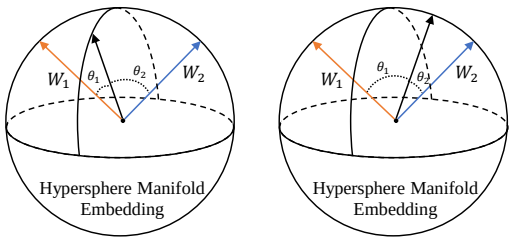
Figure 1: Heterogeneous pedestrian images. The top row is visible images captured by visible camera, bottom row is thermal images captured by thermal camera. Each column has same identity.

However, few works have paid attention to Re-ID between RGB cameras and infrared/thermal cameras, which is essentially a cross-modality problem and widely encountered in real-world scenarios. Wu *et al.* (Wu et al. 2017a) proposed a one-stream zero-padding network which learned a shared representation from heterogeneous images. Ye *et al.* (Ye et al. 2018b) proposed an end-to-end framework for visible thermal person re-identification (VT-REID), which used identity loss and ranking loss to learn a discriminative representation.

VT-REID is a very challenging problem because of the great difference between visible and thermal modalities. Visible and thermal images are intrinsically distinct. As shown in Figure 1, the first row is RGB images containing three channel information captured by visible cameras in the day, the second row is thermal images containing one channel information captured by thermal camera at night. Thus, they can be regarded as heterogeneous pedestrian images. In the view of imaging principle, the wave-length range of RGB and thermal images are different. Moreover, human pose and viewpoint change can also cause large intra-class discrepancy in VT-REID.

In this paper, we proposed an end-to-end dual-stream hy-

*Corresponding author: Nannan Wang (nnwang@xidian.edu.cn)



(a) (b)

Figure 2: 3D Sphere Softmax. W_1 and W_2 are centers of different classes, black arrow is anchor. In part (a), anchor is closer to W_1 than W_2, thus the anchor is belong to class 1. On the contrary, the anchor is classified into class 2 in part (b).

persphere manifold embedding network (HSMEnet) to learn feature representations for heterogeneous pedestrian images, which contains two separated subnets as domain-specific models. In shallower layers, the parameters of two subnets are specific for each domain to acquire domain-adaptive information. In deeper layers, shared parameters are used to learn discriminative representations for matching. We use Sphere Softmax to map the deep representation of pedestrian images onto a hypersphere. On this hypersphere, images of each identity can be classified with a clear boundary. As shown in Figure 2, classification results depend on the angle of feature vector and weight vector. And we also use Kullback Leibler (KL) Divergence to measure the matching extent of the two subnet's predictions. The main idea is that given two images of same identity from heterogeneous domains, the distributions of their predict probability should be similar. This can help the model converge to a more robust minimum value with better generalization to test data.

(a) (b) (c)

Figure 3: complex variations in heterogeneous Re-ID, (a) and (b) are different visible and thermal images of the same person, (c) is another person's images captured by visible camera. (a) shows the variations caused by human pose, (c) shows the variations caused by view points, (a)(b)(c) shows the complex variations in cross-modality matching problem.

Due to human pose and viewpoint change, another key problem for VT-REID is large intra-class discrepancy. In cross-modality matching, variations between different modalities are more complex than general matching problems as shown in Figure 3. In the view of the aforementioned two kinds of variation, we design a novel reciprocal ranking loss combined with cross-modality constraint and intra-modality constraint. For intra-modality constraint, the main idea is that the distance of anchor to cross-modality hard positive should be smaller than the anchor to intra-modality hard negative. For cross-modality constraint, the main idea is that the distance of anchor to cross-modality hard positive should smaller than the anchor to cross-modality hard negative. Since the feature vector is embedded on the hypersphere by Sphere Softmax, the distance of two features only depends on their vectorial angles.

We also propose a novel two-stage training scheme. In the first stage, the dual-stream network is trained with randomly initialized weights. In second stage, the weight matrix of Sphere Softmax is decomposed into three parts by Singular Value Decomposition(SVD). We use the product of left-unitary matrix and singular value matrix to replace the previous weight matrix as new weight matrix. Then we train the network with fixed Sphere Softmax weight matrix. As the left-unitary is orthogonal and singular value matrix is diagonal matrix, the new weight matrix is also orthogonal. Because of the orthogonal weight matrix, the deep feature representations of different person are relatively independent. Thus, the decorrelated features can achieve better performance in matching problem.

The main contributions are summarized as follows: 1) We present an end-to-end dual-stream framework for representation and metric learning, which is the first to map representation learning and metric learning both onto a hypersphere manifold. It provides a more reasonable way to combine identity loss and ranking loss. 2) We analyze the variations caused by cross-modality matching and propose a novel reciprocal ranking loss for VT-REID problem. 3) We propose a two-stage training scheme to extract decorrelated deep features of heterogeneous images.

2.Related Work

Multi-Modality Person Re-identification. Person re-identification(Re-ID) aims at spotting a particular person in other cameras. A comprehensive survey of person re-identification is provided by (Zheng, Yang, and Hauptmann 2016), so in this section, we mainly sum up the multi-modality person re-identification. Recently, a number of multi-modality person re-identification models have been proposed. Jungling *et al.* (Jungling and Arens 2010) used infrared(IR) video for Re-ID, but they only considered the IR-IR video matching. Nguyen *et al.* (Nguyen et al. 2017) firstly applied person re-identification models to visible-thermal images. Wu *et al.* (Wu, Zheng, and Lai 2017) designed a depth shape descriptor which is robust to rotation and noises. Meanwhile, Lin *et al.* (Lin et al. 2017) combined attribute information with image information for visible images. These works generally use multi-modality information to improve Re-ID performance, while we focus on cross-modality re-identification problem. For cross-modality person re-identification, Ye *et al.* (Ye et al. 2015) and Li *et al.* (Li et al. 2017b)(Li et al. 2017a) proposed a series of text-to-image person retrieval methods. However, these methods cannot be directly applied to VT-REID. In VT-REID, a two-stage framework is proposed in (Ye et al. 2018a), which

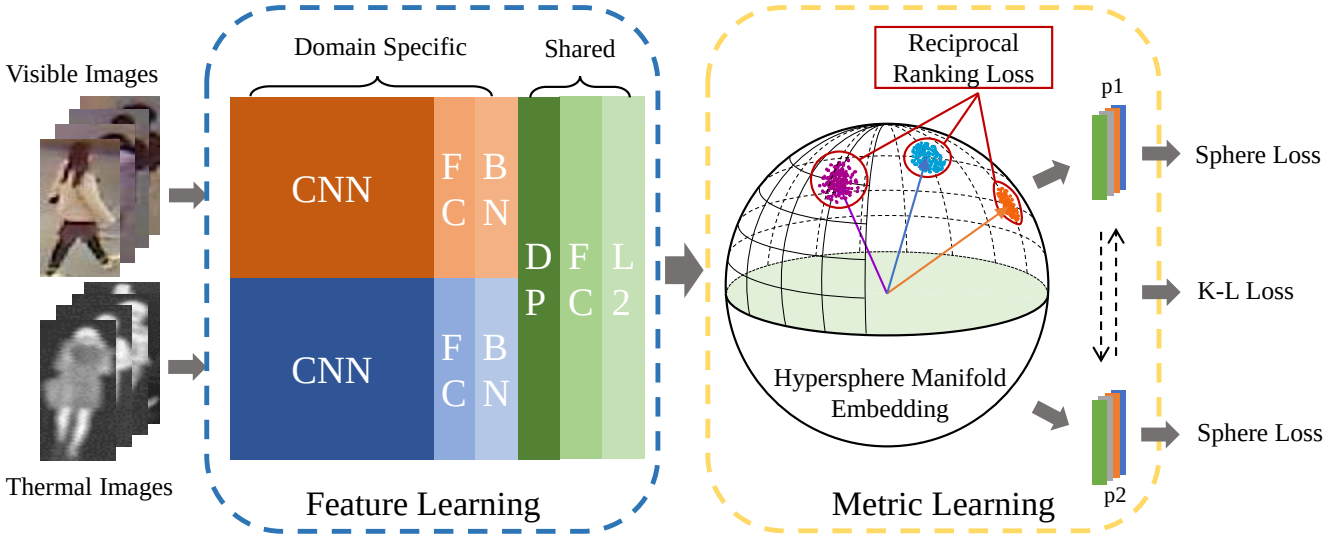


Figure 4: Pipeline of our HSME network. It includes two parts: feature learning part for extracting sharable feature representations and metric learning part for matching. The input of the network is a batch of images from visible and thermal domains. The CNN and FC are specific part to extract high-level features for pedestrian images, Orange part mean the visible specific stream and blue part represents thermal. The green part named shared layers transforms the features extracted previously into embedding space. After L2 normalization, the sharable features are mapped on a hypersphere manifold and the differences between the two samples only depend on angles. p1 and p2 are prediction probabilities of images from two domains.

contains feature learning and metric learning. In addition, Wu *et al.* (Wu et al. 2017a) introduced a deep zero-padding network to learn the shared features for different domains. In contrast, we present an end-to-end dual-stream learning framework for feature learning and metric learning.

Cross-Modality Retrieval. Cross-modality retrieval (Peng, Huang, and Zhao 2017) refers to searching instances across different modality data. It has attracted wide attention in the past few years. Representation methods include (Ding et al. 2016)(Zhang et al. 2014)(Qiu et al. 2017), these methods exhibit superior performances in many tasks. As for heterogeneous face recognition, deep feature representation learning has been considered in (He et al. 2017)(Wu et al. 2017b) for NIR-VIS face recognition. In comparison, the VT-REID task faces large intra-class variations besides cross-modality variations compared with face recognition problems, which makes these methods unsuitable for VT-REID. Ye *et al.* (Ye et al. 2018b) proposed a dual-path based network to bridge the gap between visible images and thermal images. The network contains one visible image path and one thermal image path. Under this pipeline, we design a dual-stream framework for VT-REID, which map the deep representation feature onto a hypersphere manifold.

Orthogonality in the Network. Xie *et al.* (Xie, Xiong, and Pu 2017) orthogonalized the filters of CNN and the orthogonalization improved the classification accuracy for deep networks. Sun *et al.* (Sun et al. 2017) proposed SVD-Net for person re-identification, which used Singular Vector Decomposition(SVD) to optimize the deep representation learning process. In (Sun et al. 2017), the restraint and relaxation iteration(RRI) training scheme is applied for the training process to converge. In this paper, orthogonality is used to generate decorrelated weight vectors of Sphere Softmax, so that the deep representation features that distribute on the hypersphere manifold could have less correlation as well as stronger discrimination.

3. Proposed Method

We propose an end-to-end dual-stream hypersphere manifold embedding(HSME) network for VT-REID as shown in Figure 4. The framework contains two subnets for domain specific feature learning, after extracting specific features, we use other shared weight layers to transform the features onto a common hypersphere manifold to acquire shared features. Then identity loss and ranking loss are employed to constrain the model to learn discriminative features. Furthermore, we adopt Kullback Leibler (KL) Divergence to measure the matching of predictions of two domains for better performance. Finally, to acquire low-correlated features, we modify the weight matrix of Sphere Softmax via Singular Vector Decomposition(SVD). We refer the feature correlated HSME network as D-HSME network. Our framework is divided into two parts:

Feature Learning Part. This part consists of domain specific layers and shared layers. For domain specific layers, we use two backbone networks to extract domain-specific features from heterogeneous modalities. These two backbone networks share similar structures while the parameters of each network are optimized individually. We use the AlexNet (Krizhevsky, Sutskever, and Hinton 2012) as

the backbone network and adopt image classification model pre-trained on ImageNet to boost the training process. The domain-specific features are fed into shared layers to be transformed into shared embedding space. The last layer of shared part is l2 normalization for each feature so that features are advantageous for metric learning part.

Metric Learning Part. We map the deep feature representation onto a hypersphere manifold to train the model to learn classification and identification problem on this hypersphere. We refer this process as hypersphere manifold embedding. By this way, the deep features can be more discriminative and robust for matching problem. This part is the key of HSME, which will be thoroughly illustrated from three aspects: hypersphere manifold embedding, reciprocal ranking loss and feature decorrelation.

3.1 Hypersphere Manifold Embedding

We use Sphere Softmax to map the deep features of samples onto a hypersphere manifold, so that the model can learn discriminative representations on this hypersphere. On this hypersphere, the distance between two samples can be measured by the angle of their feature vectors, which is convenient for the following metric learning process.

Revisit softmax loss. Softmax is commonly used for classification problems. For a mini-batch contains N samples, the softmax loss can be formulated as follows:

$$L_{softmax} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{W_{y_i}^T x_i}}{\sum_{j=1}^{C} e^{W_j^T x}}, \tag{1}$$

where C is the number of the classes, W is the weight matrix of classifier, x_i means the feature vector of a sample and y_i is the label of it. $W_j^T x$ means the unnormalized probabilities of $j-th$ class. Classic classification networks use a fully connected layer as softmax layer to compute probabilities, thus W is the weight of the fully connected layer. This formula can also be written as:

$$L_{softmax} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{\|W_{y_i}\|\|x_i\|\cos\theta_{y_i}}}{\sum_{j=1}^{C} e^{\|W_j\|\|x_i\|\cos\theta_j}}, \tag{2}$$

$\|W_j\|$, $\|x_i\|$ is the norm of W_j, x_i. θ_j means the angle between W_j and x_i. According to this formula, we find that both the norm and angle will influence the prediction score.

Sphere softmax loss. We fix the norm of $\|W_j\| = 1$ and $x_i = 1$ by L2 normalization as follows:

$$W_j^{'} = \frac{W_j}{\|W_j\|}, x_i^{'} = \frac{x_i}{\|x_i\|} \tag{3}$$

then we use $W_j^{'}$ and $x_i^{'}$ to replace the original W_j and x_i, the new prediction probabilities can be obtained by the following formula:

$$P_j = \|W_j\|\|x_i\|\cos\theta_j = \cos\theta_j, \tag{4}$$

then we can formulate Sphere Softmax function as:

$$L_{sphere} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s\cos\theta_{y_i}}}{\sum_{j=1}^{C} e^{s\cos\theta_j}}, \tag{5}$$

where s is the scale factor for boosting the training procedure. In this paper we empirically use s = 5 for all experiments. Additive margin softmax loss (Wang et al. 2018a) and additive angular softmax loss (Deng, Guo, and Zafeiriou 2018) are similar with formula (5), when their margin or angular is 0. But for those margin softmax losses, they need a lot of experiments to find the best combination of margin and scale factors. On the contrary, we only need to set the scale factor for the stability of training process.

For feature of any sample, the classification result only depends on the angle of feature and weight vector as shown in Figure 2. Therefore, we can map the features onto a hypersphere manifold, so that they can be discriminated by the angles. It's similar with A-Softmax loss (Liu et al. 2017a) and GA-Softmax loss (Liu et al. 2017b). But we use metric learning part to constraint the features on the manifold which is quite different with these methods.

The conventional feature embedding algorithm ends up there, but we consider that for same person from two domains, the prediction probabilities should be similar. So we use Kullback Leibler (KL) Divergence to measure the similarity of two domains' predictions $p1$ and $p2$. The KL distance from $p1$ to $p2$ is computed by:

$$D_{KL}(p_2\|p_1) = \sum_{i=1}^{N} p_2\log\frac{p_2}{p_1}, \tag{6}$$

and the overall identity loss function $L_{identity1}$ for visible stream is defined as :

$$L_{identity1} = L_{sphere1} + D_{KL}(p2\|p1), \tag{7}$$

and the loss function $L_{identity2}$ for thermal stream is:

$$L_{identity2} = L_{sphere2} + D_{KL}(p1\|p2) \tag{8}$$

3.2 Reciprocal Ranking Loss

Considering the variations from cross-modality and intra-modality, we design a novel reciprocal ranking loss for matching problem. By analyzing the samples of a mini-batch in dual-stream network, we can obtain the metric relation of a mini-batch as shown in Figure 5.

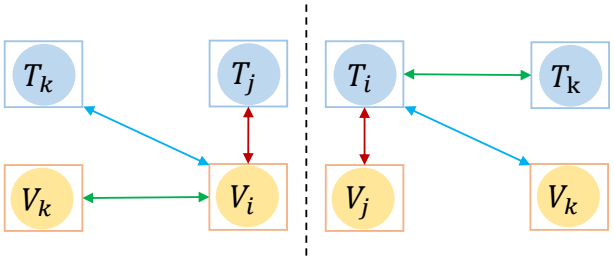


Figure 5: Samples relation in a mini-batch. Left part uses visible image as anchor and right part uses thermal image. The subscripts i and j means same identities, while i and k are different identities. V stands for visible images, T stands for thermal images.

Intra-modality constraint: This part aims at making the distance between the anchor and intra-modality hard negative larger than the distance between the anchor and cross-modality hard positive. In Figure 5, it means that the red lines should be shorter than blue ones. The constraint can be represented as follows:

$$d(v_i, t_j) < d(v_i, v_k), d(t_i, v_j) < d(t_i, t_k), \tag{9}$$

then the reciprocal ranking loss for intra-modality constraint can be formulated by:

$$\begin{aligned} L_{intra} = &\sum_{i=1}^{N} \max[(\rho + d(v_i, t_j) - \min_{\forall y_k \neq y_i} d(v_i, v_k)), 0] \\ &+ \sum_{i=1}^{N} \max[(\rho + d(t_i, v_j) - \min_{\forall y_k \neq y_i} d(t_i, t_k)), 0] \end{aligned}, \tag{10}$$

where ρ is pre-defined margin for improving the discrimination of embedding features. $d(v_i, t_j)$ represents the distance between features of V_i and T_j.

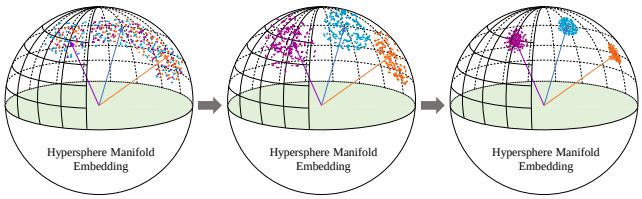


Figure 6: 3D sphere subspace:(a) samples are randomly distributed on the initialized subspace. (b) the identity loss distributes samples of each class around the weight vector of the class. (c) the final loss make the samples of each class more compact. In training stage, we will directly get the result by the final loss as shown in (c) .

Cross-modality constraint: This part aims at making the distance of anchor to cross-modality hard positive smaller than the anchor to cross-modality hard negative. In Figure 5, it means the red lines should be shorter than the blue ones. This constraint can be formulated as follows:

$$d(v_i, t_j) < d(v_i, t_k), d(t_i, v_j) < d(t_i, v_k), \tag{11}$$

like formula (10) we also use reciprocal ranking loss with pre-defined margin for cross-modality constraint:

$$\begin{aligned} L_{cross} = &\sum_{i=1}^{N} \max[(\rho + d(v_i, t_j) - \min_{\forall y_k \neq y_i} d(v_i, t_k)), 0] \\ &+ \sum_{i=1}^{N} \max[(\rho + d(t_i, v_j) - \min_{\forall y_k \neq y_i} d(t_i, v_k)), 0] \end{aligned} \tag{12}$$

For each stream, we can obtain the final loss function by combining all of the aforementioned loss functions:

$$L_{visible} = L_{identity1} + L_{intra} + L_{cross} \tag{13}$$

$$L_{thermal} = L_{identity2} + L_{intra} + L_{cross} \tag{14}$$

Training the network with $L_{visible}$ and $L_{thermal}$, we can obtain embedding features on the hypersphere manifold, and the whole procedure is illustrated on Figure 6.

3.3 Feature Decorrelation

For hypersphere space, the weight vector is randomly initialized by truncated normal distribution function. If the weight vectors of Sphere Softmax is highly correlated, the feature model learned may not be discriminative enough and the model would easily suffer from over-fitting. So we want the weight vector be less correlated and distributed on the hypersphere manifold more evenly.

In a n-dimension space, a set of orthogonal bases contain n orthogonal vectors. For softmax layer in the network, the shape of weight matrix is $m \times n (m \geq n)$, m means the embedding size of deep features and n means the class number of samples. We assume that if we can use a set of orthogonal vectors in the m-dimensional hypersphere as the weight matrix, the feature learned by the model would be more discriminative. Thus, we propose a two-stage training strategy to modify the weight vectors.

We first briefly introduce the two-stage training strategy:

Stage1: The general training procedure: We randomly initialize Sphere Softmax weight matrix by truncated normal distribution function. Then we train the model till convergence. All parameters of this model are trained.

stage2: The decorrelated training procedure: We perform SVD on the weight matrix of softmax layer as follows:

$$W = USV^T, \tag{15}$$

where W is the weight matrix of the softmax layer, U is the left-unitary matrix, S is the singular value matrix, and V is the right-unitary matrix. Then we replace the weight matrix with US. Finally, we fix the weight of softmax layer and train the network again till convergence. The optimization details are summarized in Algorithm 1.

Algorithm 1 Training Feature Decorrelated HSMEnet

Input: a HSMEnet with pre-trained weight, Re-ID training data
Output: a feature decorrelated HSMEnet(D-HSME)
stage1: train the HSMEnet until convergence
Decorrelation Decompose W with SVD decomposition, and then replace W by US
stage2: train the HSMEnet with fixed W until convergence

Considering two arbitrary weight vectors W_i, W_j on the hypersphere, which belong to i-th and j-th classes respectively, and the angle between these two vectors is θ_{ij}. f_i and f_j are two samples belong to i-th class and j-th class. After training stage 1, f_i and f_j will be constrained into the neighbors of W_i and W_j. The radius of each neighbors are small

enough, the angle between them is $\hat{\theta}_{ij} \approx \theta_{ij}$. When training stage 2 is finished, the new weight vectors of i-th and j-th classes are $W_i^{'}$ and $W_j^{'}$, and they are orthogonal, the angle between them is $\theta_{ij}^{'} = 90°$. The angle between new features which represents the two samples is close to 90°.

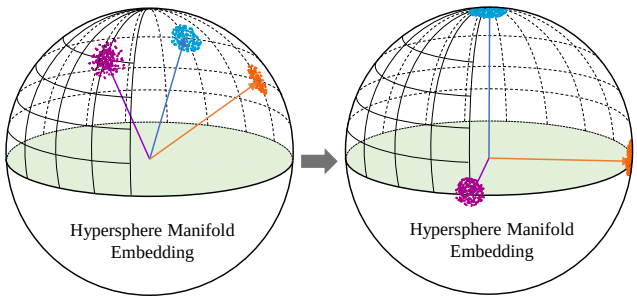


Figure 7: Feature decorrelation: (a) original weight matrix on hypersphere, (b) orthogonal weight matrix. While the weight matrix is transformed to orthogonality, the samples around different weight vector distributes with larger margin than former.

After all training stages, the weight matrix is transformed from high-correlated status to low-correlated status. The transformation is shown in Figure 7. Due to the large angle between each weight vector, the features of different samples also have larger angle than previous embedding result.

4. Experimental Results

4.1 Datasets and Evaluation

Datasets. Two public cross-modality person re-identificaition datasets are adopted to evaluate our algorithm. (Nguyen et al. 2017) provided a visible-thermal dataset RegDB, which contains 412 persons. Each person has 10 different visible and thermal images captured by visible and thermal camera. We follow the evaluation protocol in (Ye et al. 2018a), where the dataset is randomly split into training dataset and testing dataset. We repeat this split procedure for 10 trials to achieve stable results.

SYSU-MM01(Wu et al. 2017a) include RGB and infrared(IR) images of 491 identities that captured from 6 cameras, which include two IR cameras and four RGB cameras. The dataset contains 287,628 RGB images and 15,792 IR images. The training data and testing data is already splited by (Wu et al. 2017a). The training set contains 395 persons, and the testing set contains 96 persons. We adopt single-shot all-search mode evaluation protocol, since it's more challenging as mentioned in (Wu et al. 2017a).

Evaluation. For both RegDB and SYSU-MM01 dataset, multiple groundtruths exist in the gallery set. Hence, We use standard cumulated matching characteristics(CMC) curve and mean average precision(mAP) to evaluate our algorithm.

Implementation details. We use AlexNet as backbone net for both visible and thermal streams. The size of fully connected layer in shared layers is set as 1024 and the number of sample pairs in a mini-batch is set as 64 for both datasets. Dropout rate is set as 0.5. We also use random cropping for data augmentation. We set the scale factor of Sphere Loss as 5. Two Momentum optimizers are utilized for both visible and thermal streams. The artificial margin ρ is set to 0.5. The training step of stage1 is equal to stage2, which is 1000 for RegDB and 10000 for SYSU-MM01 dataset.

4.2 Comparison with State-of-the-arts

Competing algorithms. We compare HSMEnet with the state-of-the-art methods. Several other cross-modality matching methods are included for comparison. Most of the results are provided in (Ye et al. 2018b) and (Wu et al. 2017a). The competing methods include some feature learning methods(TONE, HOG, MLBP, LOMO, zero-padding, one-stream and two-stream networks(Wu et al. 2017a)). In addition, some metric learning methods are also compared, including XQDA and HCML(Jungling and Arens 2010).

Comparisons on RegDB and SYSU-MM01 are shown in Table 1. Compared with current works, HSMEnet outperforms existing state-of-the-art methods by a large margin on RegDB dataset. We report **rank-1 = 41.34%, mAP = 38.82% on RegDB, and rank-1 = 18.03%, mAP = 19.98% on SYSU-MM01.**. Meanwhile, compared with HSME, D-HSMEnet achieves significant improvement, We report **rank-1 = 50.85%, mAP = 47.00% on RegDB, and rank-1 = 20.68%, mAP = 23.12% on SYSU-MM01.**

Compared with previous best work BDTR for RegDB dataset, our D-HSMEnet improves approximately 8% $\sim$ 17% on re-identification rate and nearly 17% on mAP. For large-scale SYSU-MM01 dataset, the proposed method also achieves better performance compared with BDTR.

The advantage of our proposed method can be summarized as two folds: 1) End-to-end hypersphere manifold embedding network can extract discriminative features, which can be discriminated just by angles between them. 2) The proposed reciprocal ranking loss with dual-stream network take the complex variations for VT-REID into consideration, so that the embedding features are robust.

4.3 Ablation Study

We conduct the experiments to verify the effectiveness of the components of our framework on both RegDB and SYSU-MM01 datasets. We report the results with only identity loss, only ranking loss and HSME without KL distance as shown in Table 3.

RegDB dataset. The rank-1 accuracy is 38.15% for the ranking loss while the mAP is 30.62%. Although the CMC value is close to the result for HSME, but the mAP of ranking loss still has large margin compared with HSME. The results illustrate that identity loss can improve mAP for our model. And applying the KL distance to identity loss can improve the rank-1 accuracy and mAP value.

SYSU-MM01 dataset. The rank-1 accuracy is 13.58% for the ranking loss while the mAP is 16.63%. As shown in Table 3, ranking loss works better than identity loss. And the KL distance significantly improves the performance for all evaluation metrics.

Datasets	RegDB				SYSU-MM01			
Methods	r=1	r=10	r=20	mAP	r=1	r=10	r=20	mAP
HOG	13.49	33.22	43.66	10.31	2.76	18.25	31.91	4.24
MLBP	2.20	7.33	10.90	6.77	2.12	16.23	28.32	3.86
LOMO	0.85	2.47	4.10	2.28	1.75	14.14	26.63	3.48
One-stream	13.11	32.98	42.51	14.02	12.04	49.68	66.74	13.67
Two-stream	12.43	30.36	40.96	13.42	11.65	47.99	65.50	12.85
Zero-Padding	17.75	34.21	44.35	18.90	14.80	54.12	71.33	15.59
TONE	16.87	34.03	44.10	14.92	12.52	50.72	68.69	14.42
TONE+XQDA	21.94	45.05	55.73	21.80	14.01	52.78	68.60	14.42
TONE+HMCL	24.44	47.53	56.78	20.80	14.32	53.16	69.17	16.16
BCTR	32.67	57,64	66.58	30.99	16.12	54.90	71.47	19.15
BDTR	33.47	58.42	67.52	31.83	17.01	55.43	71.96	19.66
Ours(HSME)	41.34	65.21	75.13	38.82	18.03	58.31	74.43	19.98
Ours(D-HSME)	**50.85**	**73.36**	**81.66**	**47.00**	**20.68**	**62.74**	**77.95**	**23.12**

Table 1: Comparison with the sate-of-the-art methods on the RegDB and SYSU-MM01 datasets. Thermal images for gallery, visible images for query. CMC(%) and mAP(%)

Visible to Thermal				
Methods	r=1	r=10	r=20	mAP
TONE+HCML	24.44	47.53	56.78	20.08
Zero-Padding	17.75	56.42	67.52	31.83
BDTR	33.47	58.42	67.52	31.83
Ours(HSME)	41.34	65.21	75.13	38.82
Ours(D-HSME)	**50.85**	**73.36**	**81.66**	**47.00**
Thermal to Visible				
Methods	r=1	r=10	r=20	mAP
TONE+HCML	21.70	45.02	55.58	22.24
Zero-Padding	16.63	34.68	44.25	17.82
BDTR	32.72	57.96	68.86	31.10
Ours(HSME)	40.67	65.35	75.27	37.50
Ours(D-HSME)	**50.15**	**72.40**	**81.07**	**46.16**

Table 2: Comparison with different query settings on RegDB dataset. CMC(%) and mAP(%)

Comparing the results on both datasets, we can observe that ranking loss can get better performance than identity loss. Furthermore, KL distance can improve a lot on SYSU-MM01 than on RegDB. We assume that this is because the variations of SYSU-MM01 are more complex than RegDB, such like human pose, lighting and view point. The results of this part can verify that the combination of all different losses work best for the cross-modality person re-identification.

4.4 Different query settings

Table 2 shows the performance of different query settings as (Ye et al. 2018b). Results shows in table 3 illustrate that our method is robust to different query settings. Both HSME and D-HSME outperform the competing methods by a large gap on both settings. We assume that this advantage is caused by the symmetry of our framework.

RegDB	r=1	r=10	r=20	mAP
Only ranking loss	38.15	61.95	70.81	30.62
Only identity loss	22.02	45.47	56.26	22.42
HSME(woKL)	41.18	65.94	75.48	38.61
HSME	**41.34**	**65.21**	**75.13**	**38.82**
SYSU-MM01	r=1	r=10	r=20	mAP
Only ranking loss	13.58	54.90	72.99	16.63
Only identity loss	12.92	46.93	64.64	15.36
HSME(woKL)	16.37	56.48	74.24	18.93
HSME	**18.03**	**58.31**	**74.43**	**19.98**

Table 3: Components comparison on RegDB and SYSU-MM01 datasets. The woKL means fusion of ranking loss and identity loss without KL distance. CMC(%) and mAP(%)

Conclusion

In this paper, an end-to-end learning framework hypersphere manifold embedding(HSME) network is proposed for heterogeneous person re-identification problem. Through the HSMEnet, samples from different domains are mapped onto a hypersphere, so that features on this hypersphere can be discriminated by the angles between them. The reciprocal ranking loss designed for complex variations of cross-modality Re-ID are adopted for robust features. We further improve the HSME by decorrelating the weight matrix of the Sphere Softmax layer(D-HSME). Due to the elimination of correlation of the weight vectors, the learned embedding features suit the retrieval task better on the hypersphere manifold. Significant performance improvement is achieved on RegDB and SYSU-MM01 datasets.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 61876142, 61432014, U1605252, 61671339, 61772402, 61501339, and

in part by the National Key Research and Development Program of China under Grant 2016QY01W0200, in part by National High-Level Talents Special Support Program of China under Grant CS31117200001, in part by Young Elite Scientists Sponsorship Program by CAST (under Grant 2016QNRC001), in part by Natural Science Basic Research Plan in Shaanxi Province of China (under Grant 2017JM6085 and 2017JQ6007), in part by Young Talent fund of University Association for Science and Technology in Shaanxi, China, in part by the Fundamental Research Funds for the Central Universities under Grant XJS17086, in part by CCF-Tencent Open Fund, in part by the Xidian University-Intellifusion Joint Innovation Laboratory of Artificial Intelligence, and in part by the Innovation Fund of Xidian University.

References

Deng, J.; Guo, J.; and Zafeiriou, S. 2018. Arcface: Additive angular margin loss for deep face recognition. *arXiv preprint arXiv:1801.07698*.

Ding, G.; Guo, Y.; Zhou, J.; and Gao, Y. 2016. Large-scale cross-modality search via collective matrix factorization hashing. *IEEE Transactions on Image Processing* 25(11):5427–5440.

He, R.; Wu, X.; Sun, Z.; and Tan, T. 2017. Learning invariant deep representation for nir-vis face recognition. In *AAAI*, volume 4, 7.

Jungling, K., and Arens, M. 2010. Local feature based person reidentification in infrared image sequences. In *Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on*, 448–455. IEEE.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105.

Li, S.; Xiao, T.; Li, H.; Yang, W.; and Wang, X. 2017a. Identity-aware textual-visual matching with latent co-attention. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 1908–1917. IEEE.

Li, S.; Xiao, T.; Li, H.; Zhou, B.; Yue, D.; and Wang, X. 2017b. Person search with natural language description. *arXiv preprint arXiv:1702.05729*.

Lin, Y.; Zheng, L.; Zheng, Z.; Wu, Y.; and Yang, Y. 2017. Improving person re-identification by attribute and identity learning. *arXiv preprint arXiv:1703.07220*.

Liu, W.; Wen, Y.; Yu, Z.; Li, M.; Raj, B.; and Song, L. 2017a. Sphereface: Deep hypersphere embedding for face recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, 1.

Liu, W.; Zhang, Y.-M.; Li, X.; Yu, Z.; Dai, B.; Zhao, T.; and Song, L. 2017b. Deep hyperspherical learning. In *Advances in Neural Information Processing Systems*, 3950–3960.

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

Peng, Y.; Huang, X.; and Zhao, Y. 2017. An overview of cross-media retrieval: Concepts, methodologies, benchmarks and challenges. *IEEE Transactions on Circuits and Systems for Video Technology*.

Qiu, Z.; Pan, Y.; Yao, T.; and Mei, T. 2017. Deep semantic hashing with generative adversarial networks. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 225–234. ACM.

Sun, Y.; Zheng, L.; Deng, W.; and Wang, S. 2017. Svdnet for pedestrian retrieval. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 3820–3828. IEEE.

Wang, Z.; Hu, R.; Yu, Y.; Jiang, J.; Liang, C.; and Wang, J. 2016. Scale-adaptive low-resolution person re-identification via learning a discriminating surface. In *IJCAI*, 2669–2675.

Wang, F.; Cheng, J.; Liu, W.; and Liu, H. 2018a. Additive margin softmax for face verification. *IEEE Signal Processing Letters* 25(7):926–930.

Wang, Z.; Ye, M.; Yang, F.; Bai, X.; and Satoh, S. 2018b. Cascaded sr-gan for scale-adaptive low resolution person re-identification. In *IJCAI*, 3891–3897.

Wu, A.; Zheng, W.-S.; Yu, H.-X.; Gong, S.; and Lai, J. 2017a. Rgb-infrared cross-modality person re-identification. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 5390–5399. IEEE.

Wu, X.; Song, L.; He, R.; and Tan, T. 2017b. Coupled deep learning for heterogeneous face recognition. *arXiv preprint arXiv:1704.02450*.

Wu, A.; Zheng, W.-S.; and Lai, J.-H. 2017. Robust depth-based person re-identification. *IEEE Trans. Image Processing* 26(6):2588–2603.

Xie, D.; Xiong, J.; and Pu, S. 2017. All you need is beyond a good init: Exploring better solution for training extremely deep convolutional neural networks with orthonormality and modulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6176–6185.

Ye, M.; Liang, C.; Wang, Z.; Leng, Q.; Chen, J.; and Liu, J. 2015. Specific person retrieval via incomplete text description. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, 547–550. ACM.

Ye, M.; Lan, X.; Li, J.; and Yuen, P. C. 2018a. Hierarchical discriminative learning for visible thermal person re-identification. In *AAAI*.

Ye, M.; Wang, Z.; Lan, X.; and Yuen, P. C. 2018b. Visible thermal person re-identification via dual-constrained top-ranking. In *IJCAI*, 1092–1099.

Zhang, H.; Yang, Y.; Luan, H.; Yang, S.; and Chua, T.-S. 2014. Start from scratch: Towards automatically identifying, modeling, and naming visual attributes. In *Proceedings of the 22nd ACM international conference on Multimedia*, 187–196. ACM.

Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; and Tian, Q. 2015. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision*, 1116–1124.

Zheng, L.; Zhang, H.; Sun, S.; Chandraker, M.; Yang, Y.; Tian, Q.; et al. 2017. Person re-identification in the wild. In *CVPR*, volume 1, 2.

Zheng, L.; Yang, Y.; and Hauptmann, A. G. 2016. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*.

Zheng, Z.; Zheng, L.; and Yang, Y. 2017. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. *arXiv preprint arXiv:1701.07717* 3.