Dual-alignment Feature Embedding for Cross-modality Person Re-identification

Yi Hao
ISN State Key Lab, School of Electronic Engineering, Xidian University, Xi'an, China

Nannan Wang*
ISN State Key Lab, School of Telecommunications Engineering, Xidian University, Xi'an, China

Xinbo Gao
ISN State Key Lab, School of Electronic Engineering, Xidian University, Xi'an, China

Jie Li
ISN State Key Lab, School of Electronic Engineering, Xidian University, Xi'an, China

Xiaoyu Wang
Intellifusion, Shenzhen, China

ABSTRACT

Person re-identification aims at searching pedestrians across different cameras, which is a key problem in video surveillance. With requirements in night environment, RGB-infrared person re-identification which could be regarded as a cross-modality matching problem, has gained increasing attention in recent years. Aside from cross-modality discrepancy, RGB-infrared person re-identification also suffers from human pose and view point differences. We design a dual-alignment feature embedding method to extract discriminative modality-invariant features. The concept of dual-alignment is two folds: spatial and modality alignments. We adopt the part-level features to extract fine-grained camera-invariant information. We introduce distribution loss function and correlation loss function to align the embedding features across visible and infrared modalities. Finally, we can extract modality-invariant features with robust and rich identity embeddings for cross-modality person re-identification. Experiment confirms that the proposed baseline and improvement achieves competitive results with the state-of-the-art methods on two datasets. For instance, We achieve (57.5+12.6)% rank-1 accuracy and (57.3+11.8)% mAP on the RegDB dataset.

CCS CONCEPTS

• **Information systems** → **Top-k retrieval in databases**; **Image search**; • **Computing methodologies** → **Object identification**; **Matching**.

KEYWORDS

cross-modality, person re-identification, distribution, fine-grained

ACM Reference Format:
Yi Hao, Nannan Wang, Xinbo Gao, Jie Li, and Xiaoyu Wang. 2019. Dual-alignment Feature Embedding for Cross-modality Person Re-identification.

*Corrsponding Author: Nannan Wang(nnwang@xidian.edu.cn)

MM '19, October 21–25, 2019, Nice, France

ACM ISBN 978-1-4503-6889-6/19/10…$15.00
https://doi.org/10.1145/3343031.3351006

Figure 1: The different modality images (visible and thermal) from the real world surveillance scenario.

In *Proceedings of the 27th ACM International Conference on Multimedia (MM '19), October 21–25, 2019, Nice, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3343031.3351006

1 INTRODUCTION

Person re-identification(Re-ID) is a popular task in video surveillance, which targets to retrieve the expected pedestrian image that has the same identity with the probe image [35]. Various person Re-ID algorithms have been proposed and gained promising results [2][10][26][23][1][24][27]. In realistic scenarios, night-time or dark environment surveillance also makes up a large portion in public security. When the lighting condition is poor or unavailable, visible cameras cannot capture effective enough information and most surveillance would turn to infrared(IR) mode. Thus RGB-infrared(RGB-IR) person re-identification is introduced for robust person retrieval under low illumination environments.

As shown in Fig1, the query and gallery images are captured by different cameras. RGB camera works under good illumination environments while IR camera operates in poor ones. [25] published a large scale RGB-infrared person re-identification dataset SYSU-MM01 and addressed RGB-IR Re-ID problem with deep zero padding, which showed that RGB-IR Re-ID is challenging but still feasible. Apart from infrared images, thermal images are also used for cross-modal person Re-ID. Figure 2 shows some pedestrian images for RGB-IR Re-ID and VT-REID. In this paper, we redefine the VT-REID problem as the RGB-IR Re-ID problem since thermography is a special example of infrared imaging technology.

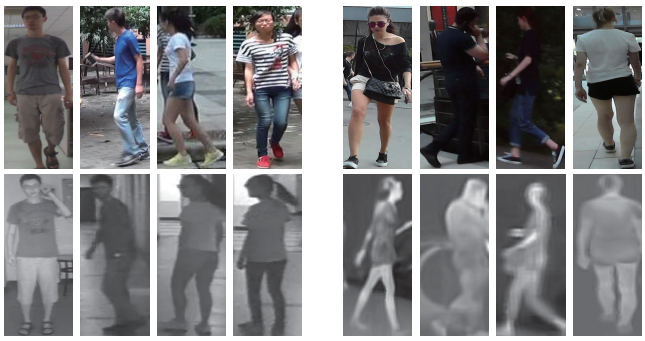

Figure 2: Samples from VT-REID and RGB-IR Re-ID. The left part is RGB and IR images, the right part is RGB and Thermal images. The discrepancy are caused from both spatial-level variants(*e.g.* human-pose, view-point) and modality-level variants(*e.g.* heterogeneity of RGB-IR images.)

The heterogeneity of cross-modality retrieval is one major challenge of RGB-IR person Re-ID problem. To address this challenge, we propose a distribution-level constraint to reduce the gap between RGB and infrared(thermal) modalities. We consider the k-channel feature map of a person image as a multivariate Gaussian distribution $N(\mu, \Sigma)$, where μ is k-dim mean vector and Σ is $k \times k$ covariance matrix. We can assume that each feature vector of the feature map is a high-dimension sample which belongs to $N(\mu, \Sigma)$. Then the distributions across different modality images should be similar for the same person. We design a distribution loss function to measure the distribution similarity between RGB and IR images. By minimizing distribution loss function, we can force the model to reduce the discrepancy between visible and infrared modalities.

Due to modality divergence, the features of different modalities are usually misalignment in correlation level. Figure 3 intuitively illustrates two correlation conditions. We use Pearson Correlation Coefficient to evaluate feature correlation in high-dimension space and design a two-stream mini-batch to input paired images into the backbone network. Each pair contains a visible and an infrared image belong to the same person. We assume that the correlations between each embedding feature in visible batch should approach those in infrared batch. We design a correlation loss function to achieve this purpose.

Aside from modality misalignment, the misalignment of human pose, viewpoint and scenarios also cause the large gap between intra-class samples. Some methods [26] used pose-based approach to extract pose-invariant features for robust feature representations. However, traditional pose estimation schemes can not obtain effective pose information for infrared pedestrian images, and some pedestrian images only contain partial human body. [22][23] made competitive results without any extra human pose information. Inspired by their methods, we also adopt a partition strategy to reduce the spatial misalignment between each pedestrian image.

The main contributions of this work are as follows:

- An end-to-end network for RGB-infrared person re-identification is introduced to align both spatial and modality inconsistency.

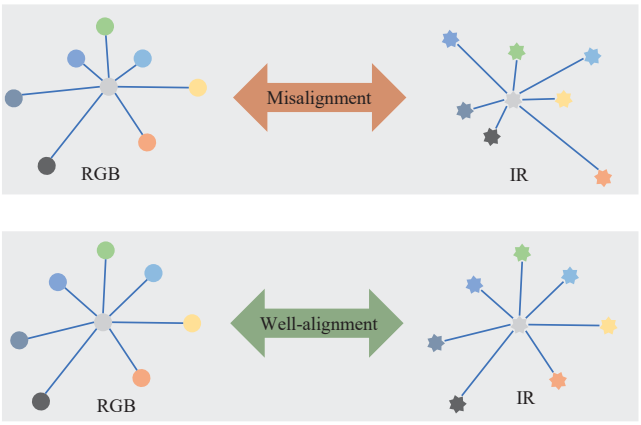

Figure 3: Two correlation conditions. Different colors represent different samples and different shapes means different modalities. The length of line between two samples represent the correlation degree, in which shorter line means higher correlation.

- We design a intra-class distribution loss function to reduce the gap between visible and infrared modalities and an intra-class correlation loss to align the feature spaces of visible images and infrared images.
- The experiment results on SYSU-MM01 and RegDB dataset show that the proposed method improves the accuracy with a large margin compared with state-of-the-art methods. Our results can be a meaningful new benchmark for RGB-IR Re-ID problem and it proves that RGB-IR Re-ID is feasible and valuable for real-world applications.

2 RELATED WORK

Person re-identification Person re-identification(Re-ID) is a popular problem in video surveillance which aims at searching a particular individual across non-overlap cameras. [35] provides a comprehensive survey of person re-identification about some early algorithms. In this section, we mainly discuss some recent algorithms which use part-level fine-grained feature representation for person Re-ID. Human pose estimation has been proved helpful to guide the network to learn pose-invariant features [19][17][14][20]. But the underlying datasets bias between pose estimation and person Re-ID makes ideal semantic segmentation and key point difficult. Moreover, only few works aim at pose estimation for infrared person image. Thus the pose-based algorithms cannot be directly applied for RGB-IR Re-ID. Other methods [22][23][21][28] abandoned external cues but also achieve competitive accuracy compared with pose-based algorithms. Therefore, we also adopt partition strategy to handle the discrimination caused by human misalignment.

Multi-modality Person Re-ID Besides human misalignment, modality shift is also a key obstacle. Previously, several multi-modal fusion models have been proposed for multi-modal person Re-ID. Some text-to-image person retrieval methods [30][9] have been proposed. [25] addressed a RGB-infrared person re-identification and design deep zero-padding method for shared features. [15] used thermal images to improve traditional visible-based Re-ID

on RegDB dataset. [29] proposed TONE to learn sharable features for VT-REID and adopted HCML to jointly optimize metrics. [31] proposed a dual-path network with bi-directional dual-constrained top-ranking loss to simultaneously handles the cross- and intra-modality variations. [3] used generative adversarial training strategy to learn discriminative feature from different modalities. [5] proposed an end-to-end dual-stream network HSME to learn feature representation for heterogeneous pedestrian images. There are also some cross-modality retrieval methods for multimodal data[33][32], which cannot be directly employed for multi-modality person Re-ID. Distinct with aforementioned methods, we consider that the deep features across modalities should follow the similar distribution. Based on this assumption, we design a distribution loss function to keep the consistency across visible and infrared modalities.

3 PROPOSED METHOD

3.1 Problem Formulation

RGB-infrared Re-ID task has two misalignment difficulties: 1) the spatial misalignment caused by human pose, view point and imperfect bounding box. 2) the modality misalignment caused by different imaging principles. We represent the multi-modal Re-ID dataset with $D = V, I$, where V denotes visible(RGB) images and I denotes infrared images. We split the dataset into training part D_{train} and testing part D_{test}. In terms of testing protocol, the RGB-IR Re-ID can be evaluated under open-set settings, which means the identities in testing set do not appear in training set.

3.2 Feature Extractor

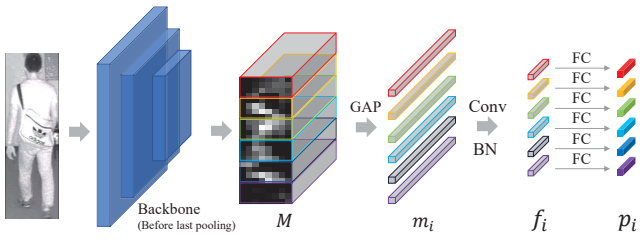

Figure 4: Structure of feature extractor. Global Average Pooling(GAP) is applied for each part instead of global feature map. Conv means 1×1 convolution operation for reducing, BN is batchnorm and FC means fully connected layers.

Figure 4 shows the structure of our base network, which use ResNet-50[6] as the backbone structure and extract the feature maps before the last pooling layer. The other advance structure for image classification can also be used without any limitations. According to [22], we slightly modify the Part-based Convolutional Baseline(PCB) for our cross-modality person re-identification task. As shown in Figure 4, we first feed input image into backbone network and obtain the multi-channel feature map M for the input image, in which M is a 3D tensor. Then we split the feature map M into K horizontal parts and apply global average pooling for each part. By this way, we get K 2048-dim vectors $m_i (i = 1, 2, ..., K)$. To reduce computation cost and parameters for fully connected layers, we employs a 1×1 Convolutional layer and a BatchNorm layer to reduce the dimension of feature vectors m_i. The m_i is transformed into 256-dim vectors, we use $f_i (i = 1, 2, ..., K)$ to represent the these 256-dim feature vectors. Attention all of the FC layers do not share weights for better performance, which has been proved in [22]. After the last set of FC layers, the base network produces the predicted label $p_i (i = 1, 2, ..., K)$ for each part. The final identity loss of base network is:

$$L_{id} = -\sum_{i=1}^{K} y_i \log(p_i), \tag{1}$$

where y_i is the ground truth label of input image.

During training stage, our base network is optimized by minimizing the sum of identity losses over K parts. After training, the output of backbone network f_i is extracted as discriminative representations for input pedestrian images with identity discerning ability. During testing stage, we concatenate all of f_i to form the final aggregated feature descriptor f_A. Different from standard PCB, we need to compute the set of distribution parameters and collect final feature descriptor f_A in training stage for modality alignment operation, which will be discussed specifically in later sections.

3.3 Intra-class Distribution Constraint

The heterogeneity of visible and infrared image is another difficulty in RGB-IR person re-identification. Previous methods often applied sample-level margin-based and some shared feature learning methods to reduce the gap between features of visible and infrared images. These methods often randomly choose positive and negative samples for anchor image across modalities, *i.e.*, Dual-Constrained Top-Ranking Loss[31] and cross-modality triplet loss[3], which are essentially a kind of large margin metric learning methods. These methods only consider the relationships among samples rather than distributions across modalities. And the performance of these algorithms are easily influenced by pre-defined margin which is usually set empirically. We assume that visible and infrared images are two forms of same information, in which they are collected by different mediums. And the neural network mainly focus on some high-level(*i.e.*, semantic-level) informations. Thus we assume that in high-level representation space, the features of visible and infrared images should follow the same distribution.

As mentioned in section 3.2, we can extract partitioned features f_i. We regard each partition as a sample point in the 256-dimension space and $f_i \sim N(\mu, \Sigma)$, where μ is 256-dim mean vector and Σ is 256×256 covariance matrix. During the training stage, we input images in a mini-batch form which include N pairs of visible and infrared paired images. As shown in Figure5, we use $f^V_{j,i} (j = 1, ..., K; i = 1, ..., N)$ to represent i-th partitioned feature of j-th visible input and $f^I_{j,i}$ for infrared input, in which $f^V_{j,i} \sim N_V(\mu^V_j, \Sigma^V_j)$ and $f^I_{j,i} \sim N_I(\mu^I_j, \Sigma^I_j)$. The superscript V and I respectively represent visible and infrared modality. For one subject, we consider that all partitioned features of visible and infrared inputs should follow the same distribution, thus we hope that $N_V(\mu^V, \Sigma^V)$ and $N_I(\mu^I, \Sigma^I)$[1] have high similarity. Inspired by

[1]For simplicity, we would omit subscript j denote $N_V(\mu^V_j, \Sigma^V_j)$ and $N_I(\mu^I_j, \Sigma^I_j)$ have the same subscript j. And for later formula, we use subscript instead superscript to represent the modality.

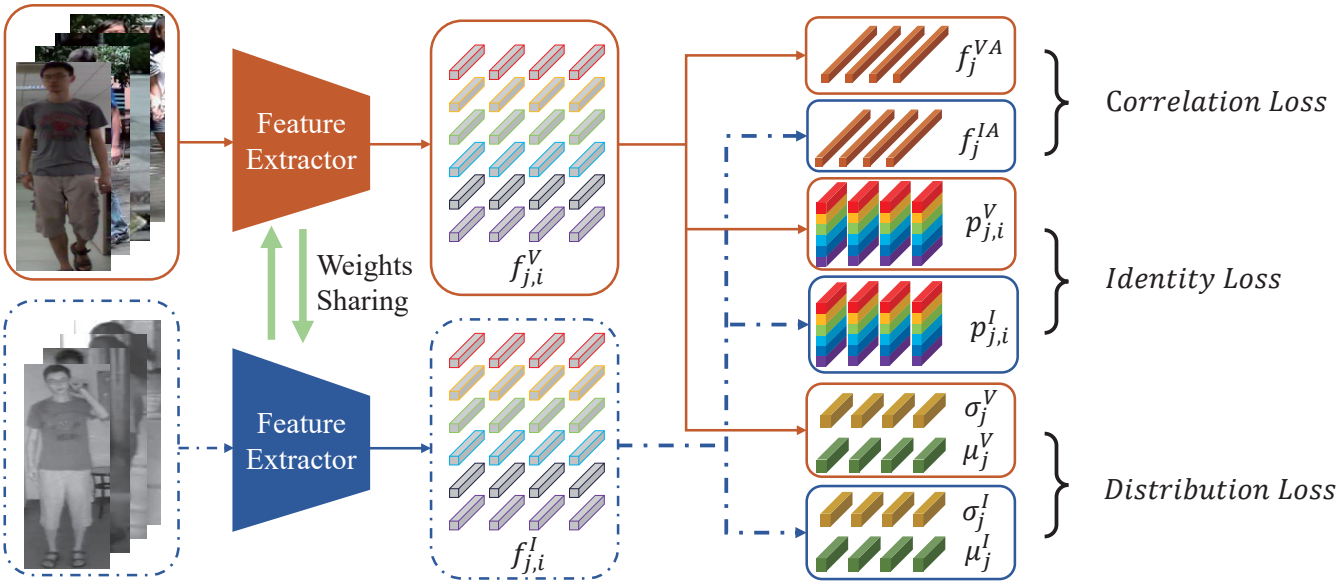

Figure 5: The framework of dual-alignment feature embedding. Orange lines and boxes represent visible data flow and blue is infrared. The details of feature extractor is shown in Figure 4. We adopt 4 pairs of RGB-IR images to illustrated the whole framework. The subscript j means the j-th($j = 1, 2, 3, 4$) samples and i means i-th partitions which is already introduced in section3.2

[4], we adopt Jensen-Shannon(JS) divergence to measure the similarity between N_V and N_I. We define the JS divergence between N_V and N_I as follows:

$$JS(N_V, N_I) = D_{KL}(N_V||N_M) + D_{KL}(N_I||N_M), \tag{2}$$

where N_M is the mixture $(N_V+N_I)/2$ and D_{KL} means the *Kullback–Leibler* (KL) divergence[8]. Given two distributions $N_0(\mu_0, \Sigma_0)$ and $N_1(\mu_1, \Sigma_1)$, in which they have the same dimension k, the KL divergence is as follows:

$$D_{KL}(N_0||N_1) = \frac{1}{2}(tr(\Sigma_1^{-1}\Sigma_0) + (\mu_1 - \mu_0)^T\Sigma_1^{-1}(\mu_1 - \mu_0) - k + \ln(\frac{\det\Sigma_1}{\det\Sigma_0})). \tag{3}$$

Because the partitioned features f_V and f_I are constrained by identity loss, thus the feature space is compact and we can use $JS_1(N_V, N_I)$[16] to measure the similarity. $JS_1(N_V, N_I)$ is also called Jeffreys divergence[18], we define the Jeffery divergence between N_I and N_V as follows:

$$JS_1(N_V, N_I) = D_{KL}(N_V||N_I) + D_{KL}(N_I||N_V), \tag{4}$$

according to equation 3, JS_1 between N_V and N_I can be written as:

$$\begin{aligned} &JS_1(N_V, N_I) \\ &= \frac{1}{2}(tr(\Sigma_I^{-1}\Sigma_V) + (\mu_I - \mu_V)^T\Sigma_I^{-1}(\mu_I - \mu_V) - k + \ln(\frac{\det\Sigma_I}{\det\Sigma_V})) \\ &+ \frac{1}{2}(tr(\Sigma_V^{-1}\Sigma_I) + (\mu_V - \mu_I)^T\Sigma_V^{-1}(\mu_V - \mu_I) - k + \ln(\frac{\det\Sigma_V}{\det\Sigma_I})). \end{aligned} \tag{5}$$

Since each channel(dim) of $f_i^V \in \mathbf{R^k}$ are relatively independent because they are extracted by K individual convolution filters, analogously f_i^I. Thus we can only consider the diagonal elements of the covariance Σ, where the other elements are zero. To minimize $JS_1(N_V, N_I)$, we can directly optimize the network by the following formula:

$$L_d = \frac{1}{2}\left[||\mu_V - \mu_I||_2^2 + ||\sigma_V - \sigma_I||_2^2\right], \tag{6}$$

where μ_V and μ_I are mean vectors of f_V and f_I, σ_V^2 and σ_I^2 are 256-dim vectors which consist of diagonal elements of covariance matrix Σ_V and Σ_I. In fact, σ can be obtained by computing the standard deviation on each dim and concatenate them as a vector. For a pair of input X_V and X_I, μ_V, μ_I, σ_V and σ_I can be computed by the following forms:

$$\mu_I = \frac{1}{K}\sum_{i=1}^{K} f_i^I, \sigma_I = \sqrt{\frac{1}{K}\sum_{i=1}^{K}\left[(f_i^I - \mu_I)\cdot(f_i^I - \mu_I)\right]}, \tag{7}$$

$$\mu_V = \frac{1}{K}\sum_{i=1}^{K} f_i^V, \sigma_V = \sqrt{\frac{1}{K}\sum_{i=1}^{K}\left[(f_i^V - \mu_V)\cdot(f_i^V - \mu_V)\right]}. \tag{8}$$

3.4 Inter-class Correlation Constraint

We design an inter-class correlation loss function to guide the model learn a consistency correlation between different modalities. The identities of visible-batch and infrared-batch is the same. Thus we consider that the images from visible-batch should have the same correlation with the infrared-batch in feature space.

We separately compute the Pearson Correlation matrix for both visible-batch and infrared-batch. We use $f_j^{VA}(j = 1, 2, ...N)$ to represent aggregated features of visible-batch images and use $f_j^{IA}(j = 1, 2, ..., N)$ to represent infrared-batch images, where N means the batch-size of each modality stream. Given two n-dim vectors

$x = [x_1, x_2, ..., x_n]^T$, $y = [y_1, y_2, ..., y_n]^T$, their Pearson correlation coefficient can be written as follows:

$$c_{xy} = \frac{\sum_{j=1}^{n}(x_j - \hat{x})(y_j - \hat{y})}{\sqrt{\sum_{j=1}^{n}(x_j - \bar{x})^2}\sqrt{\sum_{j=1}^{n}(y_j - \bar{y})^2}}, \tag{9}$$

where $\bar{x} = 1/n \sum x_i$ represents sample mean of x, and analogously for $\bar{y}$. Then we compute the Pearson correlation coefficient between each feature f_j^{VA} of visible-batch images. We can get the Pearson correlation coefficient matrix as follow:

$$C_V = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1N} \\ c_{21} & c_{22} & \cdots & c_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ c_{N1} & c_{N2} & \cdots & c_{NN} \end{bmatrix}. \tag{10}$$

For instance, c_{12} means the Pearson correlation coefficient of the 1st and the 2nd features in f^{VA}. Analogously, we can also get C_I for f^{IA}. We can simply use column vector to represent the correlation coefficient matrix C_V and C_I as follows:

$$C_V = \left[c_{V1}, c_{V2}, \ldots, c_{VN}\right]^T, C_I = \left[c_{I1}, c_{I2}, \ldots, c_{IN},\right]^T \tag{11}$$

where c_{Vj} represents the column vector which consists of the correlation coefficient between the j-th image and other images in visible-batch, analogously for c_{Ij}. Then we design the correlation loss function to minimize the difference between two correlation coefficient matrix as follows:

$$L_c = \frac{1}{N}\sum_{j=1}^{N} ||c_{Vj} - c_{Ij}||_2^2. \tag{12}$$

3.5 Feature Learning Optimization

Figure 5 shows the full framework of our proposed dual-alignment feature embedding method. The whole model can be trained in an end-to-end manner. We propose to share the weight parameters between two feature extractors, whose benefits for cross-modality Re-ID is verified by [25]. We use fusion loss function to optimize our network to address the misalignment from spatial-level and modality-level. We firstly extend formula 1 and 6 from single image and a pair image to a mini-batch condition. That is, L_{id} and L_d can be rewritten as:

$$L_{id} = -\frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{K} y_{j,i}^V \log(p_{j,i}^V) - \frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{K} y_{j,i}^I \log(p_{j,i}^I), \tag{13}$$

$$L_d = \frac{1}{2N}\sum_{j=1}^{k}(||\mu_j^V - \mu_j^I||_2^2 + ||\sigma_j^V - \sigma_j^I||_2^2), \tag{14}$$

and final loss function can be written as:

$$L = \alpha L_{id} + \beta L_d + \gamma L_c, \tag{15}$$

where α, β and γ are the trade-off factors to balance three terms. The gradient descent is adopted to optimize the network. And the last K FC layers will not be influenced when minimizing L_d and L_c. The term L_{id} will optimize the weight and bias vector of all layers in model. All of three loss functions are directly differentiable so the network can be optimized by the gradient flow.

4 EXPERIMENTS

4.1 Datasets and Protocol

RegDB dataset. RegDB dataset is collected by dual camera systems, it includes 412 identities. Each person has 10 different visible and thermal images captured by different cameras. Our experiment on this dataset follows the evaluation protocol in [31], in which the dataset is randomly halved into testing set and training set. Each subset includes 216 identities and the identities in testing set do not appear in training set. When we evaluate our model, we set 2060 thermal images for gallery images, and 2060 visible images for probe set. The procedure is repeated for 10 times, then we compute the average results to achieve statistically stable results.

SYSU-MM01 dataset. The SYSU-MM01 dataset is the first RGB-IR cross-modality Re-ID dataset, which contains 491 identities captured from four RGB cameras(1,2,4,5) and two IR cameras(3,6). The SYSU-MM01 dataset includes 287,628 RGB images and 15,792 IR images. The dataset is already separated into training set(296 identities), validation set(99 identities) and testing set(96 identities). Our experiments follows the evaluation protocol in[25]. There are two modes for comprehensively evaluating our model, *all-search* mode and *indooe-search* mode. For *all-search* mode, RGB images of testing set are used for gallery set while those IR images are used for query set. For *indoor-search* mode, RGB images of testing set captured by camera 1 and 2 are used for gallery set while IR images of testing set captured by cameras 3 are used for query set. For both modes, we adopt *single-shot* and *multi-shot* settings to compare with previous methods.

Evaluation metrics. For both RegDB and SYSU-MM01 dataset, we use standard cumulated matching characteristics(CMC) curve and mean average precision(mAP) to evaluate our algorithm. Specifically, We adopt the cumulative Matching Characteristics(CMC) at rank-1, rank-10, rank-20. For traditional visible person re-identification, there are single-query and multi-query modes in evaluation. It is worth noting that all our results are obtained in a single-query setting, which is different with multi-queries [34] setting in person re-identification.

4.2 Implementation Details

Feature extractor. We implement our algorithm with Pytorch. For RegDB dataset and SYSU-MM01 dataset, we resize the input images to 384×128, which is commonly used in deep re-ID system [21][22][7]. The ResNet model with the pre-trained parameters on ImageNet is adopted as the backbone network of our feature extractor. We set the number of parts K as 6, which is recommended in [22]. 1×1 convolutional layer is adopted to reduce 2048-dim feature vectors m_i into 256-dim. We adopt 6 separate linear layers without bias as classifiers to output the predict label of each part. Noting after 1×1 Conv layer, BatchNorm layer is used to stabilize training process and accelerate the model convergence. Different with PCB, we use the features after BatchNorm as our embedding features $f_{j,i}^V$ and $f_{j,i}^I$.

Training parameters. Our backbone uses pre-trained parameters on ImageNet. We call these parameters of backbone network as old parameters. And those parameters of 1×1 Conv layer and classifiers are called as new parameters. We use momentum optimizer to optimize both sets of parameters, the momentum is set to

0.9 and initial learning rate is respectively set to 0.01 and 0.1 for old parameters and new parameters. For both RegDB and SYSU-MM01 dataset, we set the training epoch to 20 and decay the learning rate every 5 epochs with 0.1 decay rate. The training batch-size is 32($N = 32$).

Two-stream sample. There are some difference between RegDB and SYSU-MM01 for sampling paired images. For RegDB dataset, each person has 10 visible images and 10 thermal images. We can directly randomly choose one visible image and one thermal image until traversing all samples of one person. So all images can go through the network once in each epoch. But for SYSU-MM01, there are 20284 RGB images and 9929 IR images. Thus, most of IR images go through the network more than one times in each epoch.

Trade-off factors. The identity loss is designed to make features discriminative while the distribution loss and correlation loss are addressed for learning modality-invariant features. We think that discrimination and modality-invariant have equal effects on the cross-modality retrieval results. Thus we set three trade-off factors α, β and γ by 1:1:1 proportion. Noting for RegDB dataset, we find distribution loss and correlation loss converge slowly while identity loss decreases quickly. We think this is because that the RegDB dataset is small-scale and the variations for each individual are less than variations between different modalities. Meanwhile, the discrepancy of RGB-Thermal is bigger than RGB-IR, which also caused the slow convergence of distribution loss and correlation loss. Thus we reset the trade-off α, β and γ by 1:11:2 proportion for RegDB dataset, which is best combination from our multiple cross-validation. We also provide the results which are conducted by 1:1:1 proportion on RegDB dataset in ablation study part and they also achieve competitive improvement.

4.3 Ablation Study

Variants evaluation. We report the evaluations of the proposed end-to-end dual-alignment feature embedding method(DFE) with different variants. The results on RegDB dataset are shown in Figure 6. Here, "baseline" means the results which conducted only with identity loss L_{id}. "DFE w/o l_c" means the network are trained by fusion loss which combine the identity loss l_{id} and distribution loss l_d. "DFE w/o l_d" means the fusion loss consists of identity loss l_{id} and correlation l_c. "DFE" expresses the performance with all three terms l_{id}, l_d and l_c in fusion loss. "DFE o.p." represents the results of DFE when trade-off factors is original 1:1:1 proportion.

Experiment results shown in Figure 6 illustrate that intra-class distribution constraint could definitively improve the performance of baseline method by 6-7% on Rank-1 accuracy and mAP. It verifies the idea that extracting modality-invariant features by closing the distributions of one subject across two modalities can help to handle large discrepancy caused by modality variations. According to Figure 6, we can see that the correlation loss l_c also achieve competitive improvement by aligning correlation between two-modalities. Furthermore, all CMC and mAP are further improved by integrating all modality-level alignment strategies l_d and l_c. Overall, the proposed DFE improves the rank-1 accuracy from 57.5% to 70.13%, and mAP from 57.3% to 69.14% on RegDB dataset.

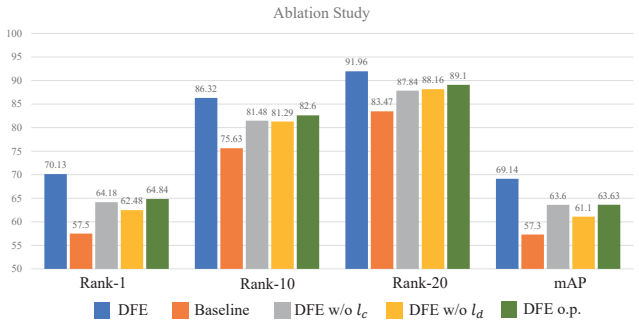

Figure 6: Evaluations of different variants of the proposed dual-alignment embedding method on the RegDB dataset. CMC(%) and mAP(%).

4.4 Results of SYSU-MM01 Dataset

The results of SYSU-MM01 dataset are shown in Table 1. There are two competing methods focus on RGB-IR Re-ID. Deep Zero-Padding [25] utilized a one-stream network to acquire shared information across domains. The cmGAN [3] showed excellent performance on SYSU-MM01 dataset. In table 1, we show the results of these state-of-the-art methods on SYSU-MM01, including rank-1, rank-10, rank-20 and mAP. Moreover, we also exhibit the results of three general deep models: one-stream, two-stream, and asymmetric FC. The detail description of these three models can be found in [25]. Most of the results are originated from [3] and [25] on the SYSU-MM01 dataset.

In Table 1, the results of two rows on the bottom are conducted by our baseline and improvement with dual-alignment embedding. It shows that the idea of combining with intra-class distribution constraint and inter-class correlation constrain contribute to the final retrieval results. So our dual-alignment embedding network can do better in reducing discrepancy caused by modality variations. As shown in Table 1, our dual-alignment embedding algorithm undoubtedly outperforms all existing methods in all protocols. Specifically, the proposed DFE outperforms the previous best methods cmGAN and it's variations on all search single-shot mode in terms of rank-1 accuracy by 21.74%(48.71-26.97) and mAP by 20.79%(48.59-27.80). Our significant results are meaningful for RGB-IR Re-ID, which proves that RGB-IR Re-ID is feasible can be handled with our advance dual-alignment embedding method. Moreover, our improvement with correlation loss l_c and distribution loss l_d outperforms our baseline method in terms of every metrics, which also illustrates that the proposed distribution loss and correlation loss are effective and robust to different retrieval scenario. As shown in Table 1, the performance of DFE w/o l_c is slightly lower than DFE on each metrics, which means the combination of all different losses works best for the cross-modality person re-identification on SYSU-MM01 dataset. And the results of DFE w/o l_c are simultaneously higher than baseline can also achieve the effectiveness of distribution loss l_d.

In addition, we also compare our proposed method with some advance visible-thermal re-identification (VT-REID) methods on SYSU-MM01 dataset, the experiment results are shown in Table 2. We adopt the *all-search* single-shot mode evaluation protocol. A

Method	all-search								indoor-search							
	single-shot				multi-shot				single-shot				multi-shot			
	r1	r10	r20	mAP	r1	r10	r20	mAP	r1	r10	r20	mAP	r1	r10	r20	mAP
HOG+Euclidean	2.76	18.25	31.91	4.24	3.82	22.77	37.63	2.16	3.22	24.68	44.52	7.25	4.75	29.06	49.38	3.51
HOG+CRAFT	2.59	17.93	31.50	4.24	3.58	22.90	38.59	2.06	3.03	24.07	42.89	7.07	4.16	27.75	47.16	3.17
HOG+CCA	2.74	18.91	32.51	4.28	3.25	21.82	36.51	2.04	4.38	29.96	50.43	8.70	4.62	34.22	56.28	3.87
HOG+LFDA	2.33	18.58	33.38	4.35	3.82	20.48	35.84	2.20	2.44	24.13	45.50	6.87	3.42	25.27	45.11	3.19
LOMO+CCA	2.42	18.22	32.45	4.19	2.63	19.68	34.82	2.15	4.11	30.60	52.54	8.83	4.86	34.40	57.30	4.47
LOMO+CRAFT	2.34	18.70	32.93	4.22	3.03	21.70	37.05	2.13	3.89	27.55	48.16	8.37	2.45	20.20	38.15	2.69
LOMO+CDFE	3.64	23.18	37.28	4.53	4.70	28.23	43.05	2.28	5.75	34.35	54.90	10.19	7.36	40.38	60.44	5.64
LOMO+LFDA	2.98	21.11	35.36	4.81	3.86	24.01	40.54	2.61	4.81	32.16	52.50	9.56	6.27	36.29	58.11	5.15
Asymmetric FC	9.30	43.26	60.38	10.82	13.06	52.11	69.52	6.68	14.59	57.94	78.68	20.33	20.09	69.37	85.08	13.04
Two-stream	11.65	47.99	65.50	12.85	16.33	58.35	74.46	8.03	15.60	61.18	81.02	21.49	22.49	72.22	88.61	13.92
One-stream	12.04	49.68	66.74	13.67	16.26	58.14	75.05	8.59	16.94	63.55	82.10	22.95	22.62	71.74	87.82	15.04
Zero-padding	14.80	54.12	71.33	15.95	19.13	61.40	78.41	10.89	20.58	68.38	85.79	26.92	24.43	75.86	91.32	18.64
cmGAN	26.97	67.51	80.56	27.80	31.49	72.74	85.01	22.27	31.63	77.23	89.18	42.19	37.00	80.94	92.11	32.76
Ours(baseline)	43.13	85.37	93.67	43.56	48.69	88.94	95.63	36.92	45.24	86.40	94.75	53.75	50.38	92.02	96.98	44.48
Ours(DFE w/o l_c)	45.81	86.81	94.59	46.37	52.26	90.58	96.14	39.76	51.32	89.17	96.46	58.73	57.97	92.42	97.6	49.66
Ours(DFE)	**48.71**	**88.86**	**95.27**	**48.59**	**54.63**	**91.62**	**96.83**	**42.14**	**52.25**	**89.86**	**95.85**	**59.68**	**59.62**	**94.45**	**98.07**	**50.60**

Table 1: Comparison with the sate-of-the-art methods on the SYSU-MM01 datasets. We use infrared image to search visible image. CMC(%) and mAP(%)

Method	Rank 1	Rank 10	Rank 20	mAP
TONE	12.52	50.72	68.69	14.42
TONE+XQDA	14.01	52.78	68.60	14.42
TONE+HMCL	14.32	53.16	69.17	16.16
TONE+MLAPG	12.43	50.64	68.72	14.61
Top-Ranking	12.96	51.80	71.00	16.11
BCTR	16.12	54.90	71.47	19.15
BDTR	17.01	55.43	71.96	19.66
HSME	18.03	58.31	74.43	19.98
D-HSME	20.68	62.74	77.95	23.12
Ours(baseline)	43.13	85.37	93.67	43.56
Ours(DFE)	**48.71**	**88.86**	**95.27**	**48.59**

Table 2: Comparison with the VT-REID methods on the SYSU-MM01 datasets in all-search single-shot mode. CMC(%) and mAP(%)

series of TONE methods[29]firstly address VT-REID problem. BDTR [31] and it's variants BCTR, which are the previous competitive works on visible-thermal person re-identification. HSME [5] and its improvement D-HSME are state-of-the-art methods for VT-REID. The results are shown in Table 2.

To intuitively illustrate the effectiveness of our correlation loss function, we choose two sets of features and plot their correlation coefficient values as shown in Figure 7. In Figure 7, the top set is extracted by the proposed DFE while the bottom set is extracted by "DFE w/o l_c". The value of each point in line is the correlation coefficient between the 0-th features and other 31 features. The correlation of the top set is consistent across modalities, as opposed to the bottom set. Figure 8 shows top-10 ranking results for some infrared query images, which illustrate great robustness for pose and illuminations of our proposed DFE.

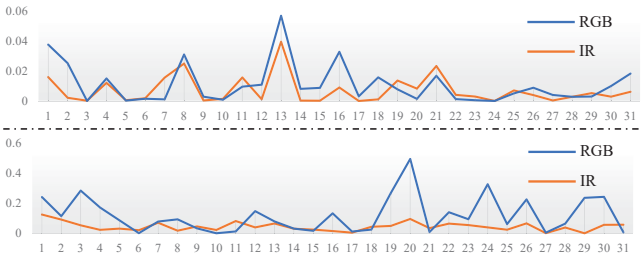

Figure 7: Correlation coefficient of two set of features from RGB-IR paired image. The value of correlation coefficient is scaled into [0,2] for better display.

4.5 Results of RegDB Dataset

In this section, we compare our proposed method and its variations with the state-of-the-art VT-REID and RGB-IR Re-ID methods on RegDB dataset. Several other cross-modality matching methods are also included for comparison. Most of the results are provided in[31]. The competing algorithms contains some feature-learning methods(HOG, LOMO[11], TONE[29], zero-padding, one-stream, two-stream[25]). Moreover, we include some metric learning methods for comparison, including XQDA[12], GSM[13], MLAPG[12] and HCML[29].

The results of RegDB dataset are shown in Table 3. Compared with current state-of-the-art methods, DFE consistently outperforms them with more than 20% for all evaluation index. Specifically, we achieve rank-1=70.13%, rank-10=86.32%, rank-20=91.96% and mAP=69.14%. The advantages of the proposed DFE include two folds: 1) part-based fine-grained feature learning method can extract spatial-alignment features with discriminative information. 2) intra-class distribution loss and inter-class correlation loss make the extracted feature more modality-invariant for cross-modality person re-identification.

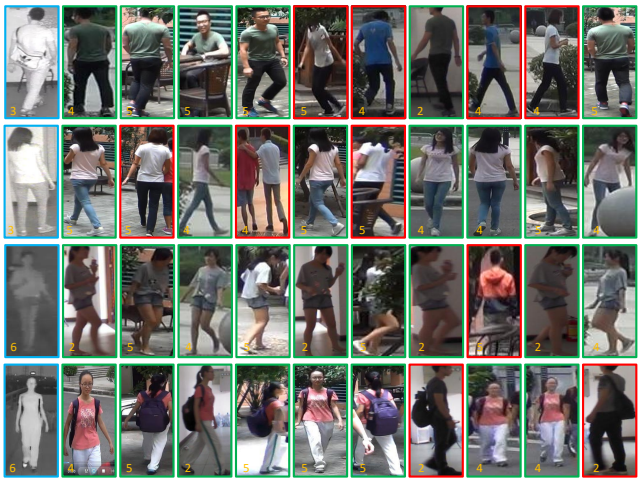

Figure 8: Top-10 ranking list of some infrared query images on SYSU-MM01 dataset. The images with green borders belongs to the same identity as the given query which has blue border, red is opposite. Yellow number at the left corner of each image is camera number.

Method	Rank 1	Rank 10	Rank 20	mAP
HOG	13.49	33.22	43.66	10.31
GSM	17.28	34.47	45.26	15.06
LOMO	0.85	2.47	4.10	2.28
One-stream	13.11	32.98	42.51	14.02
Two-stream	12.43	30.36	40.96	13.42
Zero-Padding	17.75	34.21	44.35	18.90
TONE+XDQA	21.94	45.05	55.73	21.80
TONE+MLAPG	17.82	40.29	49.73	18.03
TONE+HMCL	24.44	47.53	56.78	20.80
BCTR	32.67	57.64	66.58	30.99
BDTR	33.47	58.42	67.52	31.83
HSME	41.34	65.21	75.13	38.82
D-HSME	50.85	73.36	81.66	47.00
Ours(baseline)	57.50	75.63	83.47	57.30
Ours(DFE w/o L_c)	64.18	81.48	87.84	63.60
Ours(DFE w/o L_d)	62.48	81.29	88.16	61.10
Ours(DFE)	**70.13**	**86.32**	**91.96**	**69.14**

Table 3: Comparison with the sate-of-the-art methods on the RegDB datasets. We use visible image to search thermal image. CMC(%) and mAP(%)

Different query settings. We also conduct the experiment of different query settings on the RegDb dataset, which is also an evaluation protocol in [31]. The experiment results shown in table 4 illustrate that our method is robust to different query settings. The performance of visible-to-thermal is close to the results of thermal-to-visible with less than 3% difference of Rank-1 accuracy and mAP, which demonstrates that our proposed method is flexible and applicable in realistic applications.

Visible to Thermal				
Methods	Rank 1	Rank 10	Rank 20	mAP
TONE+HCML	24.44	47.53	56.78	20.80
Zero-Padding	17.75	56.42	67.52	31.83
BDTR	33.47	58.42	67.52	31.83
HSME	41.34	65.21	75.13	38.82
D-HSME	50.85	73.36	81.66	47.00
Ours(baseline)	57.50	75.63	83.47	57.30
Ours(DFE)	**70.13**	**86.32**	**91.96**	**69.14**
Thermal to Visible				
Methods	Rank 1	Rank 10	Rank 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
HSME	40.67	65.35	75.27	37.50
D-HSME	50.15	72.40	81.07	46.16
Ours(baseline)	53.20	70.20	77.43	53.1
Ours(DFE)	**67.99**	**85.56**	**91.41**	**66.70**

Table 4: Comparison with different query settings on RegDB dataset. CMC(%) and mAP(%)

4.6 Conclusions

In this paper, we propose an end-to-end learning framework via dual-alignment embedding method for cross-modality person re-identification. The proposed method aligns the inconsistency on both spatial-level and modality-level. To address the misalignment from spatial-level, we use part-based CNN model to learn discriminative part-level deep features. Meanwhile, to address the large discrepancy between visible images and infrared images, we design an intra-class distribution loss function and inter-class correlation loss function to learn modality-invariant features. Extensive experiments on two common cross-modality person re-identification datasets illustrate the superiority and robustness of our proposed method when compared with the state-of-the-arts.

5 ACKNOWLEDGE

This work was supported in part by the National Natural Science Foundation of China (under Grant 61876142, 61432014, U1605252, 61772402, and 61671339), in part by the National Key Research and Development Program of China under Grant 2016QY01W0200, in part by the National High-Level Talents Special Support Program of China under Grant CS31117200001, in part by the Young Elite Scientists Sponsorship Program by CAST under Grant 2016QNRC001, in part by the Young Talent fund of University Association for Science and Technology in Shaanxi, China, in part by the CCF-Tencent Open Research Fund (No. RAGR20180105) and Tencent AI Lab Rhino-Bird Focused Research Program (No. JR201923), and in part by the Xidian University-Intellifusion Joint Innovation Laboratory of Artificial Intelligence, in part by the Fundamental Research Funds for the Central Universities under Grant JB190117, and in part by the Innovation Fund of Xidian University.

REFERENCES

[1] Susanne Boll, Kyoung Mu Lee, Jiebo Luo, Wenwu Zhu, Hyeran Byun, Chang Wen Chen, Rainer Lienhart, and Tao Mei (Eds.). 2018. *2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, October 22-26, 2018*. ACM. https://doi.org/10.1145/3240508

[2] Ying-Cong Chen, Xiatian Zhu, Wei-Shi Zheng, and Jian-Huang Lai. 2018. Person re-identification by camera correlation aware feature augmentation. *IEEE transactions on pattern analysis and machine intelligence* 40, 2 (2018), 392–408.

[3] Pingyang Dai, Rongrong Ji, Haibin Wang, Qiong Wu, and Yuyu Huang. 2018. Cross-Modality Person Re-Identification with Generative Adversarial Training.. In *IJCAI*. 677–683.

[4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*. 2672–2680.

[5] Yi Hao, Nannan Wang, Jie Li, and Xinbo Gao. 2019. HSME: Hypersphere Manifold Embedding for Visible Thermal Person Re-Identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 8385–8392.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[7] Houjing Huang, Wenjie Yang, Xiaotang Chen, Xin Zhao, Kaiqi Huang, Jinbin Lin, Guan Huang, and Dalong Du. 2018. EANet: Enhancing Alignment for Cross-Domain Person Re-identification. *arXiv preprint arXiv:1812.11369* (2018).

[8] Solomon Kullback. 1997. *Information theory and statistics*. Courier Corporation.

[9] Shuang Li, Tong Xiao, Hongsheng Li, Wei Yang, and Xiaogang Wang. 2017. Identity-aware textual-visual matching with latent co-attention. In *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE, 1908–1917.

[10] Wei Li, Xiatian Zhu, and Shaogang Gong. 2018. Harmonious attention network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2285–2294.

[11] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. 2015. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2197–2206.

[12] Shengcai Liao and Stan Z Li. 2015. Efficient psd constrained asymmetric metric learning for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*. 3685–3693.

[13] Liang Lin, Guangrun Wang, Wangmeng Zuo, Xiangchu Feng, and Lei Zhang. 2017. Cross-domain visual matching via generalized similarity measure and feature learning. *IEEE transactions on pattern analysis and machine intelligence* 39, 6 (2017), 1089–1102.

[14] Jinxian Liu, Bingbing Ni, Yichao Yan, Peng Zhou, Shuo Cheng, and Jianguo Hu. 2018. Pose Transferrable Person Re-Identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[15] Dat Tien Nguyen, Hyung Gil Hong, Ki Wan Kim, and Kang Ryoung Park. 2017. Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors* 17, 3 (2017), 605.

[16] Frank Nielsen. 2010. A family of statistical symmetric divergences based on Jensen's inequality. *arXiv preprint arXiv:1009.4004* (2010).

[17] Xuelin Qian, Yanwei Fu, Tao Xiang, Wenxuan Wang, Jie Qiu, Yang Wu, Yu-Gang Jiang, and Xiangyang Xue. 2018. Pose-Normalized Image Generation for Person Re-identification. In *The European Conference on Computer Vision (ECCV)*.

[18] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. 2000. The earth mover's distance as a metric for image retrieval. *International journal of computer vision* 40, 2 (2000), 99–121.

[19] M Saquib Sarfraz, Arne Schumann, Andreas Eberle, and Rainer Stiefelhagen. 2018. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 420–429.

[20] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. 2018. Mask-Guided Contrastive Attention Model for Person Re-Identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[21] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. 2018. Part-aligned bilinear representations for person re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 402–419.

[22] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. 2018. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European Conference on Computer Vision (ECCV)*. 480–496.

[23] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. 2018. Learning discriminative features with multiple granularities for person re-identification. In *2018 ACM Multimedia Conference on Multimedia Conference*. ACM, 274–282.

[24] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. 2018. Learning Discriminative Features with Multiple Granularities for Person Re-Identification, See [1], 274–282. https://doi.org/10.1145/3240508.3240552

[25] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. 2017. RGB-Infrared Cross-Modality Person Re-identification. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 5390–5399.

[26] Jing Xu, Rui Zhao, Feng Zhu, Huaming Wang, and Wanli Ouyang. 2018. Attention-aware compositional network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2119–2128.

[27] Jiwei Yang, Xu Shen, Xinmei Tian, Houqiang Li, Jianqiang Huang, and Xian-Sheng Hua. 2018. Local Convolutional Neural Networks for Person Re-Identification, See [1], 1074–1082. https://doi.org/10.1145/3240508.3240645

[28] Hantao Yao, Shiliang Zhang, Richang Hong, Yongdong Zhang, Changsheng Xu, and Qi Tian. 2019. Deep representation learning with part loss for person re-identification. *IEEE Transactions on Image Processing* (2019).

[29] Mang Ye, Xiangyuan Lan, Jiawei Li, and Pong C Yuen. 2018. Hierarchical discriminative learning for visible thermal person re-identification. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[30] Mang Ye, Chao Liang, Zheng Wang, Qingming Leng, Jun Chen, and Jun Liu. 2015. Specific person retrieval via incomplete text description. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*. ACM, 547–550.

[31] Mang Ye, Zheng Wang, Xiangyuan Lan, and Pong C Yuen. 2018. Visible Thermal Person Re-Identification via Dual-Constrained Top-Ranking.. In *IJCAI*. 1092–1099.

[32] Yi Yu, Suhua Tang, Kiyoharu Aizawa, and Akiko Aizawa. 2018. Category-based deep CCA for fine-grained venue discovery from multimodal data. *IEEE transactions on neural networks and learning systems* 30, 4 (2018), 1250–1258.

[33] Yi Yu, Suhua Tang, Francisco Raposo, and Lei Chen. 2019. Deep cross-modal correlation learning for audio and lyrics in music retrieval. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 15, 1 (2019), 20.

[34] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. 2015. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision*. 1116–1124.

[35] Liang Zheng, Yi Yang, and Alexander G Hauptmann. 2016. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984* (2016).