Cross-Modality Earth Mover's Distance for Visible Thermal Person Re-identification

Yongguo Ling[1], Zhun Zhong[2], Zhiming Luo[1], Fengxiang Yang[1], Donglin Cao[1*], Yaojin Lin[3], Shaozi Li[1], Nicu Sebe[2]

[1]Department of Artificial Intelligence, Xiamen University, China
[2]Department of Information Engineering and Computer Science, University of Trento, Italy
[3]School of Computer Science, Minnan Normal University, China
yongguo_ling@sina.com {zhiming.luo, yangfx.stu, another, szlig}@xmu.edu.cn
{zhun.zhong, niculae.sebe}@unitn.it yjlin@mnnu.edu.cn

Abstract

Visible thermal person re-identification (VT-ReID) suffers from inter-modality discrepancy and intra-identity variations. Distribution alignment is a popular solution for VT-ReID, however, it is usually restricted to the influence of the intra-identity variations. In this paper, we propose the Cross-Modality Earth Mover's Distance (CM-EMD) that can alleviate the impact of the intra-identity variations during modality alignment. CM-EMD selects an optimal transport strategy and assigns high weights to pairs that have a smaller intra-identity variation. In this manner, the model will focus on reducing the inter-modality discrepancy while paying less attention to intra-identity variations, leading to a more effective modality alignment. Moreover, we introduce two techniques to improve the advantage of CM-EMD. First, Cross-Modality Discrimination Learning (CM-DL) is designed to overcome the discrimination degradation problem caused by modality alignment. By reducing the ratio between intra-identity and inter-identity variances, CM-DL leads the model to learn more discriminative representations. Second, we construct the Multi-Granularity Structure (MGS), enabling us to align modalities from both coarse- and fine-grained levels with the proposed CM-EMD. Extensive experiments show the benefits of the proposed CM-EMD and its auxiliary techniques (CM-DL and MGS). Our method achieves state-of-the-art performance on two VT-ReID benchmarks.

Introduction

Person re-identification (ReID) is critical in a safety surveillance system, aiming at matching a query of interest from a set of gallery images captured by non-overlapping cameras. Traditional ReID (Lin et al. 2019; Wu, Zhu, and Gong 2020; Zheng et al. 2021) assumes that the images are collected by RGB cameras. However, the appearance characteristics of visible (RGB) images are largely relying on the illumination conditions and will be highly influenced by the poor illumination (*e.g.,* night-time). To handle the night-time context, the ReID community has captured night-time images by thermal cameras and proposed a task called visible thermal person re-identification (VT-ReID) (Ye et al. 2019). This task is a cross-modality retrieval problem, aiming to search a query of one modality (*e.g.*, visible) from the gallery of another modality (*e.g.*, thermal). VT-ReID is more challenging than the traditional ReID, because we need to jointly overcome the inter-modality discrepancy caused by visible and thermal cameras and the intra-identity variations caused by different factors (such as view, pose and background).

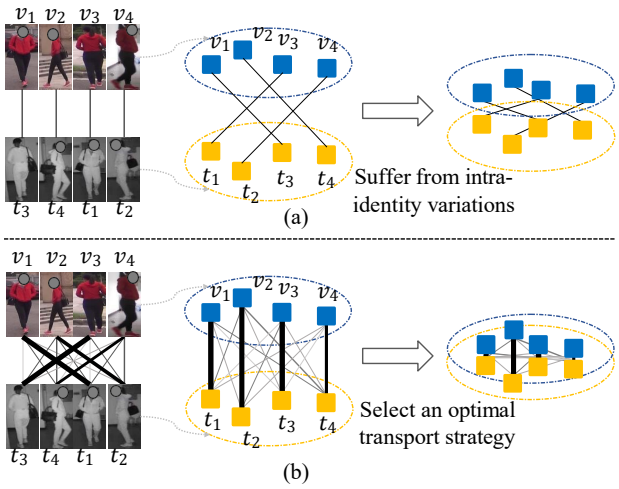

Figure 1: (a) Traditional distribution alignment. The randomly selected pairs usually have high intra-identity variations, hindering the reduction of inter-modality discrepancy. (b) Distribution alignment with Cross-Modality Earth Mover's Distance (CM-EMD). Our CM-EMD assigns large weights to pairs that have less intra-identity variations, leading the model to focus on reducing the modality discrepancy rather than intra-identity variations.

Distribution alignment (Pu et al. 2020; Hao et al. 2019a; Dai et al. 2018) is a popular and effective solution for VT-ReID, which aims to reduce the distribution gap between visible and thermal modalities. However, the existence of intra-identity variations may interfere with the optimization of distribution alignment. For example, given a training mini-batch, the randomly selected cross-modality pairs of

*Corresponding author

the same identity commonly have a large variation in pose, view and background (Fig. 1 (a)). In this context, when the cross-modality discrepancy is reduced to a certain extent and is not dominant, the model will turn to decrease the intra-identity variations that are not suitable for optimizing by the distribution alignment function. Intuitively, if there are only cross-modality differences between all the selected cross-modality pairs, the model can always focus on reducing the modality gap to obtain a better modality alignment result. However, it is hard to achieve the above situation, since, in VT-ReID, intra-identity variations always exist and such pairs are hard to obtain using only the identity annotations.

In this paper, we propose a novel distribution alignment approach for VT-ReID, called Cross-Modality Earth Mover's Distance (CM-EMD), which can largely mitigate the impact of the intra-identity variations during modality alignment. Specifically, given a mini-batch, CM-EMD leverages EMD (Rubner, Tomasi, and Guibas 1998) to select an optimal transport strategy across two modalities and assigns a large transport weight between two cross-modality samples with a smaller intra-identity variation (Fig. 1 (b)). In this way, the model can focus on reducing the modality discrepancy instead of the intra-identity variation, leading to an effective modality alignment. We also introduce two techniques to facilitate the effectiveness of CM-EMD. (1) The Cross-Modality Discrimination Learning (CM-DL). As CM-EMD mainly focuses on aligning the cross-modality distribution, it will gently close the distance of cross-modality negative pairs (Fig. 3 (a$\rightarrow$c)), degrading the discrimination ability of the representation. Therefore, we propose CM-DL to improve the discrimination by reducing the ratio between the intra-identity and the inter-identity variances. (2) The Multi-Granularity Structure (MGS). As local information is important for re-ID (Sun et al. 2018; Wang et al. 2018; Wei et al. 2021a; Zhang et al. 2021a), we introduce MGS to perform finer modality alignment. In our MGS, we first extract both part-based local features and global features and then apply CM-EMD to them individually, enabling us to decrease the modality gap on both coarse-grained and fine-grained feature levels. In summary, the contributions of this paper are:

- We introduce the Cross-Modality Earth Mover's Distance (CM-EMD) for VT-ReID, which can alleviate the negative effect caused by the intra-identity variations and effectively bridge the modality gap.
- We propose the Cross-Modality Discrimination Learning (CM-DL), effectively overcoming the discrimination degradation problem raised by CM-EMD.
- We design the Multi-Granularity Structure (MGS), versatilely reducing the modality gap with CM-EMD.

Experiments on two VT-ReID datasets verify the advantages of the proposed CM-EMD, CM-DL and MGS, and demonstrate the superior performance of our method over state-of-the-art methods.

Related Work

VT-ReID has been first introduced by (Wu et al. 2017), which aims to match the query person of one modality from the gallery of another one. Since then, many methods were proposed for VT-ReID, which can be mainly divided into four groups. **(1) Feature extractor based methods** (Wu et al. 2017; Ye et al. 2018b, 2020b; Liu et al. 2020; Yang et al. 2020; Lu et al. 2020; Ye et al. 2020c; Jia et al. 2020; Chen et al. 2021; Fu et al. 2021; Tian et al. 2021) mainly focus on designing a cross-modality network to extract modality-invariant and discriminative representation. **(2) Metric learning based methods** (Ye et al. 2018a; Hao et al. 2019b; Ye et al. 2020a; Feng, Lai, and Xie 2019; Liu, Tan, and Zhou 2020; Ling et al. 2020, 2021) are proposed to learn an embedding space by explicitly enforcing the intra-class samples of two modalities close to each other. **(3) Distribution alignment based methods** (Dai et al. 2018; Hao et al. 2019a; Pu et al. 2020; Zhang et al. 2021b; Wu et al. 2021; Park et al. 2021) are mainly proposed to decrease the distribution divergence to smooth the inter-modality discrepancy and gain modality invariant feature. **(4) Image generation based methods** (Wang et al. 2019b,a; Li et al. 2020; Wei et al. 2021b; Choi et al. 2020; Wang et al. 2020; Liu et al. 2021) commonly attempt to generate fake/virtual images that are used to bridge the modality gap in the image-level.

Earth Mover's Distance (EMD) (Rubner, Tomasi, and Guibas 1998) is a metric to estimate the distance between two distributions, which is a special case of the transportation problem from linear optimization. EMD was applied to address many tasks, such as image retrieval (Rubner, Tomasi, and Guibas 2000), tracking (Zhao, Yang, and Tao 2008; Li 2013; Schulter et al. 2017), graph matching (Nikolentzos, Meladianos, and Vazirgiannis 2017), document retrieval (Chen et al. 2019a), overcoming mode collapsing (Chen et al. 2018; Salimans et al. 2018; Arjovsky, Chintala, and Bottou 2017) of Generative Adversarial Network (GAN) (Goodfellow et al. 2014), and few-shot learning (Zhang et al. 2020). Different from them, this work employs EMD to solve the problem of VT-ReID, where we propose the cross-modality EMD to effectively learn modality-invariant representation.

Methodology

Overview. The framework of the proposed method is shown in Fig. 2. We first extract global features, local features, and accumulated local features with the proposed multi-granularity structure (MGS) for both visible and thermal modalities. We then calculate the losses of identity classification and losses of the proposed Cross-Modality Earth Mover's Distance (CM-EMD) with the three types of features. For the proposed Cross-Modality Discrimination Learning (CM-DL), we first obtain a holistic feature by concatenating the weighted local features and computing the loss of CM-DL with the holistic feature. The loss of identity classification aims to learn a basic representation. The loss of CM-EMD is designed to effectively reduce the modality discrepancy. The loss of CM-DL focuses on learning more discriminative representation.

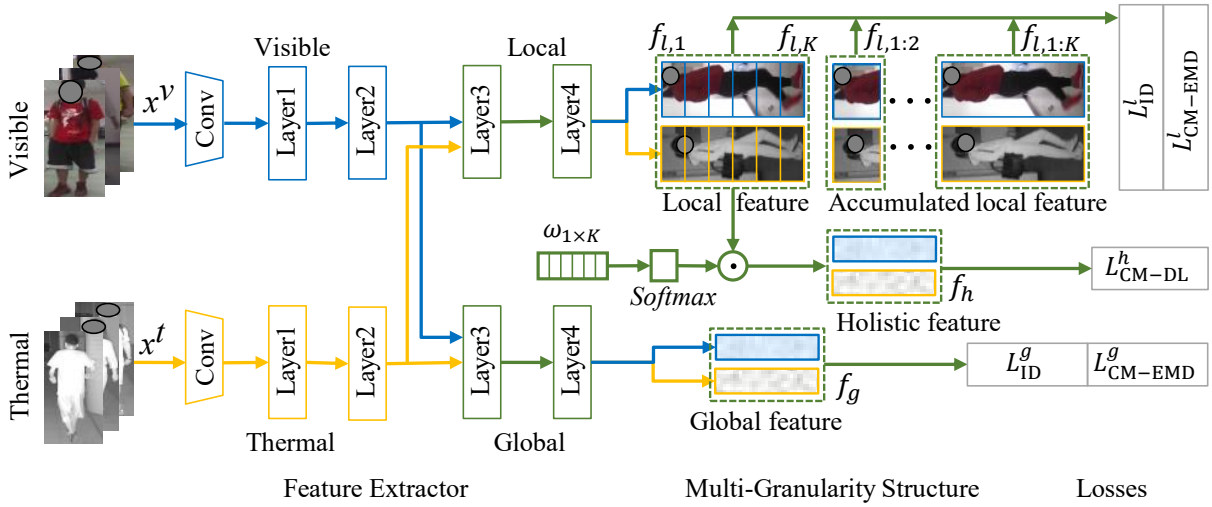

Figure 2: The framework of the proposed method. In the shallow layers, we use different parameters for visible modality and thermal modality. The shallow layers are the first convolutional layer, 1-th residual block, and 2-th residual block of the ResNet-50 (He et al. 2016). After the shallow layers, we design two streams for extracting global-based features and local-based features. Each stream includes the 3-th and 4-th residual blocks of the ResNet-50, which are shared by both modalities. The parameters of the two streams are different. Given the inputs, we first extract global features, local features, and accumulated local features with our Multi-Granularity Structure (MGS). We then calculate the losses of classification and the losses of Cross-Modality Earth Mover's Distance with the obtained three types of features. On the other hand, we generate the holistic feature by concatenating the weighted local features. The loss of Cross-Modality Discrimination Learning is calculated with the holistic feature.

Basic Loss

Identity classification loss is widely used in ReID, which is obtained by calculating the cross-entropy loss with the identity labels. The identity classification loss is referred to L_{ID} in this paper. Because the data in VT-ReID come from both visible and thermal modalities, we need to reduce the modality discrepancy during training so that the model can generate discriminative cross-modality representation. Next, we introduce a novel distribution alignment method to achieve the above goal.

Cross-Modality Earth Mover's Distance

Motivation. Distribution alignment is an effective way to reduce the modality discrepancy for VT-ReID. However, as discussed in the introduction, distribution alignment will be affected by the intra-identity variations. In our intuition, if we can select cross-modality pairs that have fewer intra-identity variations, the modality alignment process can mainly focus on reducing the modality discrepancies caused by the selected pairs. However, the above situation is hard to achieve since the intra-identity variations are caused by various factors while we only have the identity information. On the other hand, EMD (Rubner, Tomasi, and Guibas 1998) is a measure of the distance between two distributions, which can be solved by minimizing the cost of transporting one distribution to another. As a result, the two samples that are more similar will have a high connected weight otherwise have a low connected weight. Inspired by this, we introduce a new modality alignment method, named Cross-Modality Earth Mover's Distance (CM-EMD). Taking advantage of EMD, CM-EMD can automatically assign weights between samples according to their similarities. As a consequence, the cost of CM-EMD is largely dominated by the pairs that have fewer intra-identity variations (*i.e.*, have high similarities) and the impact of the intra-identity variations can be largely suppressed. We next introduce CM-EMD in detail.

Given a training mini-batch, we have N^v visible modality samples and N^t thermal modality samples, which are randomly selected from C identities. The features obtained by the model are defined as $\mathbb{F}^v$ and $\mathbb{F}^t$ for visible modality and thermal modality, respectively. We will introduce how to obtain the representation of samples in Sec. . The feature distributions of visible modality and thermal modality are denoted as $\boldsymbol{\nu} \in \mathbf{P}(\mathbb{F}^v)$ and $\boldsymbol{\tau} \in \mathbf{P}(\mathbb{F}^t)$, respectively. $\Pi(\boldsymbol{\nu}, \boldsymbol{\tau})$ represents all joint distributions $\rho(\boldsymbol{f^v}, \boldsymbol{f^t})$, where $\boldsymbol{f^v} \in \mathbb{F}^v$ and $\boldsymbol{f^t} \in \mathbb{F}^t$. The CM-EMD among the samples of two modalities can be defined as:

$$\mathcal{D}_{\text{CM-EMD}}(\mathbb{F}^v, \mathbb{F}^t) = \inf_{\boldsymbol{\rho} \in \Pi(\boldsymbol{\nu}, \boldsymbol{\tau})} \mathbb{E}_{(\boldsymbol{f^v}, \boldsymbol{f^t}) \sim \boldsymbol{\rho}}[M(\boldsymbol{f^v}, \boldsymbol{f^t})], \quad (1)$$

where $M(\boldsymbol{f^v}, \boldsymbol{f^t})$ is the cost function and is calculated by the euclidean distance between $\boldsymbol{f^v}$ and $\boldsymbol{f^t}$. EMD has the form of transportation problem (TP) from Linear Programming, and Eq. 1 can be re-formulated as:

$$\mathcal{D}_{\text{CM-EMD}}(\mathbb{F}^v, \mathbb{F}^t) = \min_{\mathbf{S} \in \Pi(\mathbf{V}, \mathbf{T})} \sum_{i=1}^{N^v} \sum_{j=1}^{N^t} \mathbf{S}_{ij} \cdot M\left(\boldsymbol{f_i^v}, \boldsymbol{f_j^t}\right), \quad (2)$$

where $\mathbf{V} = \{v_i\}_{i=1}^{N^v}$ and $\mathbf{T} = \{t_i\}_{i=1}^{N^t}$ are N^v-dim simplex and N^t-dim simplex, respectively. v_i and t_j are the weights

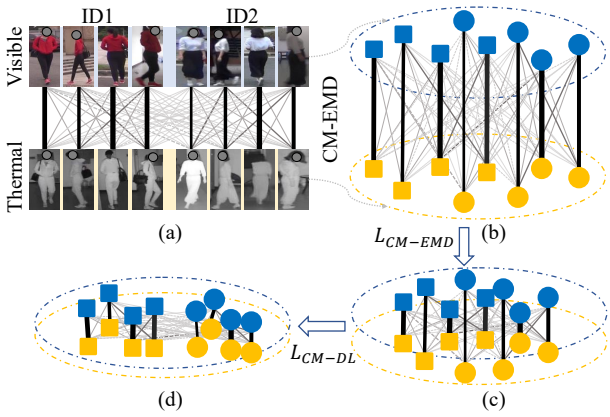

Figure 3: Illustration of the proposed CM-EMD (a→c) and CM-DL (c→d). CM-EMD can effectively reduce the modality gap while CM-DL can effectively promote the discrimination ability of the model. The pairs with high weights are connected with wider lines. Shapes indicate the identities. Colors represent the modalities (blue for visible and yellow for thermal).

of their corresponding nodes. $\Pi(\mathbf{V}, \mathbf{T})$ denotes all transport plans $\mathbf{S} \in \mathbb{R}_{+}^{N^v \times N^t}$. S_{ij} indicates the weight of shifting from v_i to t_j, which subjects to:

$$S_{ij} \geqslant 0, \quad \sum_{j=1}^{N^t} S_{ij} = v_i, \quad \sum_{i=1}^{N^v} S_{ij} = t_j, \\ \forall i = 1, \ldots, N^v, \quad \forall j = 1, \ldots, N^t. \tag{3}$$

To reduce the computational complexity, we use Sinkhorn algorithm (Cuturi 2013; Peyré, Cuturi et al. 2019) to solve the Eq. 2.

Discussion. After optimization, there is a negative correlation between the transport weight S_{ij} and the transport cost $M(\boldsymbol{f_i^v}, \boldsymbol{f_j^t})$. The two samples ($f_i^v$ and f_j^t) will be assigned with a large transport weight when they are close to each other, otherwise will be assigned with a small transport weight. As a result, when using Eq. 2 as the loss function of modality alignment, the model will mainly focus on reducing the discrepancy of samples that have fewer intra-identity variations (*i.e.*, have small transport cost). Meanwhile, the samples that have large intra-identity variations will have less impact on the model optimization, since they are assigned with low weights. An example of CM-EMD is illustrated in Fig. 3 (a→c).

Cross-Modality Discrimination Learning

One disadvantage of modality alignment is that the optimization process may degrade the discrimination capability of the model (Chen et al. 2019b), which is also a problem of the proposed CM-EMD. For example, all cross-modality pairs are considered since CM-EMD mainly concentrates on modality alignment. It will slightly increase the similarity of cross-modality negative pairs (Fig. 3 (a→c)). Class-wise variance can well represent the identity distributions. Specifically, if the model can well distinguish samples of different identities, the intra-class variance should be small while the inter-class variance should be large. Taking this into mind, we propose the Cross-Modality Discrimination Learning (CM-DL) to improve the discrimination ability of the model, which is achieved by constraining the relation between cross-modality intra-class variance and cross-modality inter-class variance.

Given a training mini-batch, we first calculate the feature mean for each modality, formulated as:

$$\mu^v = \frac{1}{N^v} \sum_{i=1}^{N^v} f_i^v, \quad \mu^t = \frac{1}{N^t} \sum_{i=1}^{N^t} f_i^t. \tag{4}$$

We then calculate the class-wise feature mean for each modality, formulated as:

$$\mu_c^v = \frac{1}{N_c^v} \sum_{i=1}^{N_c^v} f_{i,c}^v, \quad \mu_c^t = \frac{1}{N_c^t} \sum_{i=1}^{N_c^t} f_{i,c}^t, \tag{5}$$

where N_c^v and N_c^t denote the number of samples of class c for visible modality and thermal modality, respectively. $f_{i,c}$ is the feature of the sample that belongs to the class c.

The cross-modality intra-class variance is defined as:

$$V_{intra}(\mathbb{F}^v, \mathbb{F}^t) = \sum_{c=1}^{C} \left\{ \sum_{i=1}^{N_c^t} \left(f_{i,c}^t - \mu_c^v\right) \left(f_{i,c}^t - \mu_c^v\right)^\top \right. \\ \left. + \sum_{i=1}^{N_c^v} \left(f_{i,c}^v - \mu_c^t\right) \left(f_{i,c}^v - \mu_c^t\right)^\top \right\}, \tag{6}$$

where $(\cdot)^\top$ represents the transpose operation. Similarly, the cross-modality inter-class variance is defined as:

$$V_{inter}(\mathbb{F}^v, \mathbb{F}^t) = \sum_{c=1}^{C} \left\{ N_c^v \left(\mu_c^v - \mu^t\right) \left(\mu_c^v - \mu^t\right)^\top \right. \\ \left. + N_c^t \left(\mu_c^t - \mu^v\right) \left(\mu_c^t - \mu^v\right)^\top \right\}. \tag{7}$$

By jointly considering the cross-modality intra-class variance and cross-modality inter-class variance, our CM-DL can be formulated as:

$$L_{\text{CM-DL}} = \frac{V_{intra}(\mathbb{F}^v, \mathbb{F}^t)}{V_{inter}(\mathbb{F}^v, \mathbb{F}^t)}. \tag{8}$$

By minimizing Eq. 8, the model is encouraged to reduce the cross-modality intra-class variance as well as increase the cross-modality inter-class variance, leading to a more discriminative representation. An example of CM-DL is shown in Fig. 3 (c→d).

Multi-Granularity Structure

Some works (Wei et al. 2021a; Zhang et al. 2021a) have demonstrated the effectiveness of local features. Inspired by this, we introduce the multi-granularity structure (MGS) to align the modalities in both coarse- and fine-grained aspects. Given the global and local features of the last residual convolution layer $f_{g,res}$ and $f_{l,res}$, we extract three types of

features: global features, local features, accumulated local features.

Global feature. We directly apply the generalized-mean pooling (GeM) (Radenović, Tolias, and Chum 2018) and the batch normalization neck (BNNeck) (Luo et al. 2019) on $f_{g,res}$ and obtain the global feature f_g.

Local feature. We first evenly divide $f_{l,res}$ into K parts along the vertical direction and perform the GeM and BNNeck on them, obtaining K local features, $\{f_{l,1}, ..., f_{l,K}\}$.

Accumulated local feature. To obtain more diverse features, we extract accumulated local features by combining the different numbers of part features, which are represented by $\{f_{l,1:2}, ..., f_{l,1:K}\}$. For example, $f_{l,1:2}$ is obtained by concatenating the local features $f_{l,1}$ and $f_{l,2}$.

Losses for CM-EMD. Given the multi-granularity features, the losses of CM-EMD can be calculated as:

$$\begin{aligned} L^g_{\mathrm{CM-EMD}} &= D_{\mathrm{CM-EMD}}\left(\mathbb{F}^v_g, \mathbb{F}^t_g\right), \\ L^l_{\mathrm{CM-EMD}} &= \sum_{k=1}^{K} D_{\mathrm{CM-EMD}}\left(\mathbb{F}^v_{l,k}, \mathbb{F}^t_{l,k}\right) \\ &\quad + \alpha \sum_{k=2}^{K} D_{\mathrm{CM-EMD}}\left(\mathbb{F}^v_{l,1:k}, \mathbb{F}^t_{l,1:k}\right), \end{aligned} \tag{9}$$

where α is the hyper-parameter and K is the part number. $L^g_{\mathrm{CM-EMD}}$ is the global-based loss of CM-EMD and $L^l_{\mathrm{CM-EMD}}$ is the local-based loss of CM-EMD.

Loss for CM-DL. CM-DL is mainly used to alleviate the discrimination degradation problem raised by CM-EMD. To reduce the computational cost of CM-DL, instead of using the global-based and local-based features individually, we concatenate the weighted local features to produce a holistic feature, which is utilized for the calculated loss of CM-DL. Specifically, the holistic feature is represented as:

$$f_h = [\omega_1 f_{l,1} \mid \omega_2 f_{l,2} \mid \cdots \mid \omega_K f_{l,K}], \tag{10}$$

where $[\cdot \mid \cdot]$ indicates the concatenation operation. $\omega_1, ..., \omega_K$ are the trainable weights, which are normalized by the SoftMax function. The loss of CM-DL is reformulated as follows:

$$L^h_{\mathrm{CM-DL}} = \frac{V_{intra}(\mathbb{F}^v_h, \mathbb{F}^t_h)}{V_{inter}(\mathbb{F}^v_h, \mathbb{F}^t_h)}. \tag{11}$$

Losses for identity classification. As explained in Sec. , L_{ID} is calculated by identity classification loss. We apply L_{ID} on the global feature, local features, and accumulated local features. Specifically, we add classification heads (1 for global features, K for local features, and K for accumulated local features), which are fully-connected layers with an output length of #identities, after these three types of features. The losses of identity classification can be calculated as:

$$\begin{aligned} L^g_{\mathrm{ID}} &= \mathrm{CrossEntropy}\left(\mathbb{H}_g, \mathbb{Y}\right), \\ L^l_{\mathrm{ID}} &= \sum_{k=1}^{K} \mathrm{CrossEntropy}\left(\mathbb{H}_{l,k}, \mathbb{Y}\right) \\ &\quad + \alpha \sum_{k=2}^{K} \mathrm{CrossEntropy}\left(\mathbb{H}_{l,1:k}, \mathbb{Y}\right), \end{aligned} \tag{12}$$

where $\mathbb{H}$ indicates the predictions of the corresponding classification heads and $\mathbb{Y}$ is the identity labels.

Overall

Training. By considering the proposed CM-EMD, CM-DL, MSG and the basic loss, the model is optimized by:

$$\begin{aligned} \arg\min_{\theta} \gamma_1 L^h_{\mathrm{CM-DL}} + \gamma_2 L^l_{\mathrm{ID}} + \gamma_3 L^l_{\mathrm{CM-EMD}} \\ + \gamma_4 L^g_{\mathrm{ID}} + \gamma_5 L^g_{\mathrm{CM-EMD}}, \end{aligned} \tag{13}$$

where $\gamma_{1\to5}$ are the hyper-parameters that balance the importance of different losses.

Testing. In the testing phase, we obtain the final feature by concatenating the global feature and local features:

$$f_{test} = [\beta f_{l,1:K} \mid (1-\beta) f_g], \tag{14}$$

where β is the hyper-parameter that controls the importance of global and local features.

Experiment

Experimental Settings

Datasets. Experiments are conducted on two VT-ReID datasets. *SYSU-MM01* (Wu et al. 2017) are captured by four RGB cameras and two thermal cameras, respectively. The training set contains 22,258 RGB images and 11,909 infrared images of 395 identities. The testing set involves 3,803 query (infrared) images and 301 gallery (RGB) images of 96 identities. For evaluation, we use two testing modes, *i.e.,* all-search mode and indoor-search mode, we report single-shot settings for these two modes. *RegDB* (Nguyen et al. 2017) comprises 4,120 RGB images and 4,120 infrared images of 412 identities, collected from one RGB camera and one infrared camera. Each identity has 10 RGB images and 10 infrared images. For evaluation, we equally divide RegDB into the training and testing sets. Two testing settings are used: Thermal (query) to Visible (gallery) setting and Visible (query) to Thermal (gallery) setting.

Evaluation metrics. The Cumulative Matching Characteristics (CMC) and mean Average Precision (mAP) are used to evaluate the retrieval performance. For CMC, we report the rank-1 (R1) accuracy.

Implementation details. The baseline model based on (Ye et al. 2021b). We use random cropping for data augmentation during training. For each training mini-batch, we set the number of identities to 6 (*i.e.*, C=6) for both datasets. We then randomly sample 8 RGB images and 8 infrared images (*i.e.*, $N^v_c = N^t_c = 8$) for SYSU-MM01, and 4 RGB images and 4 infrared images (*i.e.*, $N^v_c = N^t_c = 4$) for RegDB. The input images are resized to $384 \times 192 \times 3$. The SGD optimizer is used for training and the initial learning rate is set to 0.1 with a warm-up strategy (Luo et al. 2019). We divide the learning rate by 10 after 20 and 50 epochs. We train the model for a total of 80 epochs. For the hyper-parameters of the proposed model, the features are divided into 6 parts (i.e. $K = 6$). we set α (Eq. 9) to 0.2 and 1.0 for SYSU-MM01 and RegDB, respectively. $\gamma_{1\to5}$ (Eq. 13) are to $\{1, 1, 0.1, 2, 0.1\}$ for SYSU-MM01 and $\{3, 2, 0.4, 1, 0.6\}$ for RegDB, respectively. During testing, we set β (Eq. 14) to 0.7 and 0.5 for SYSU-MM01 and RegDB, respectively.

Datasets		RegDB				SYSU-MM01			
Settings		Visible to Thermal		Thermal to Visible		All-search		Indoor-search	
Method	Venue	R1	mAP	R1	mAP	R1	mAP	R1	mAP
Zero-Padding (Wu et al. 2017)	ICCV17	17.75	18.90	/	/	14.80	15.95	20.58	26.92
HCML (Ye et al. 2018a)	AAAI18	24.44	20.8	21.70	22.24	14.32	16.16	/	/
D2RL (Wang et al. 2019b)	CVPR19	/	/	43.4	44.1	28.9	29.2	/	/
D-HSME (Hao et al. 2019b)	AAAI19	50.85	47	50.15	46.16	20.68	23.12	/	/
AlignGAN (Wang et al. 2019a)	ICCV19	57.9	53.6	56.3	53.4	42.4	40.7	45.9	54.3
Hi-CMD (Choi et al. 2020)	CVPR20	/	/	70.93	66.04	34.94	35.94	/	/
PIG (Wang et al. 2020)	AAAI20	48.50	49.3	48.1	48.90	38.1	36.9	43.8	52.9
CMM-CML (Ling et al. 2020)	ACMMM20	59.81	60.86	/	/	51.8	51.21	54.98	63.7
Xmodal (Li et al. 2020)	AAAI20	62.21	60.18	/	/	49.92	50.73	/	/
DDAG (Ye et al. 2020c)	ECCV20	69.34	63.46	68.06	61.80	54.75	53.02	61.02	67.98
cm-SSFT (Lu et al. 2020)	CVPR20	73.3	72.9	71.0	71.7	61.6	63.2	70.5	72.6
CMAlign (Park et al. 2021)	ICCV21	74.17	67.64	72.43	65.46	55.41	54.14	58.46	66.33
NFS (Chen et al. 2021)	CVPR21	80.54	72.1	77.95	69.79	56.91	55.45	62.79	69.79
VCD-VML (Tian et al. 2021)	CVPR21	73.2	71.6	71.8	70.1	60.02	58.80	66.05	72.98
MCLNet (Hao et al. 2021)	ICCV21	80.3	73.07	75.93	69.49	65.40	61.98	72.56	76.58
MPANet (Wu et al. 2021)	CVPR21	83.7	80.9	82.8	80.7	70.58	68.24	76.74	80.95
SMCL (Wei et al. 2021b)	ICCV21	83.93	79.83	83.05	78.57	67.39	61.78	68.84	75.56
CM-NAS (Fu et al. 2021)	ICCV21	84.54	80.32	82.57	78.31	61.99	60.02	67.01	72.95
CAJL (Ye et al. 2021a)	ICCV21	85.03	79.14	84.75	77.82	69.88	66.89	76.26	80.37
DART (Yang et al. 2022)	CVPR22	83.60	75.67	81.97	73.78	68.72	66.29	72.52	78.17
MID (Huang et al. 2022)	AAAI22	87.45	84.85	84.29	81.41	60.27	59.40	64.86	70.12
FMCNet (Zhang et al. 2022)	CVPR22	89.84	84.43	88.38	83.86	66.34	62.51	68.15	74.09
Ours	AAAI23	**94.37**	**88.23**	**92.77**	**86.85**	**73.39**	**68.56**	**80.53**	**82.71**

Table 1: Comparison with the state-of-the-art methods on the RegDB and SYSU-MM01 datasets.

Comparison with the State of the Art

To demonstrate the superiority of our method, we compare it against the state-of-the-art approaches on SYSU-MM01 and RegDB datasets in Table 1.

Results on RegDB. We first compare our method with state-of-the-art methods on RegDB. It is clear that our method outperforms the state-of-the-art methods by a large margin in both evaluation settings. Specifically, we obtain **rank-1 = 94.37%** and **mAP = 88.23%** for the "Visible to Thermal" setting, and, **rank-1 = 92.77%** and **mAP = 86.85%** for "Thermal to Visible" setting. Compared to the current best competitor (FMCNet (Zhang et al. 2022) in CVPR 2022), we show significant improvements. For example, our method is higher than FMCNet by 4.53% in rank-1 accuracy and 3.80% in mAP on the "Visible to Thermal" setting, and, by 4.39% in rank-1 accuracy and 2.99% in mAP on the "Thermal to Visible" setting, respectively. This indicates that our method produces a new state-of-the-art performance on RegDB.

Results on SYSU-MM001. For the results on SYSU-MM001, our method also achieves the best results in all settings. Specifically, we achieve **rank-1 = 73.39%** and **mAP = 68.56%** for the all-search mode, and, **rank-1 = 80.53%** and **mAP = 82.71%** for the indoor-search mode. Compared to the current best competitor (MPANet (Wu et al. 2021) published in CVPR 2021), our method clearly surpasses MPANet by 2.81% and 3.79% in rank-1 accuracy on the two evaluation settings, respectively.

#	Method	SYSU-MM01		RegDB	
		R1	mAP	R1	mAP
0	Baseline w/ Global	54.22	51.18	46.46	43.48
1	+CM-EMD	**60.01**	**54.75**	**76.76**	**69.70**
2	Baseline w/ Local	55.56	53.00	64.85	61.90
3	+CM-EMD	65.97	62.34	89.37	81.19
4	+CM-EMD+CM-DL	**67.81**	**63.68**	**93.45**	**85.33**
5	Baseline w/ MGF	61.58	59.26	69.08	64.22
6	+CM-EMD	71.26	66.59	92.86	84.99
7	+CM-EMD+CM-DL	**73.39**	**68.56**	**94.37**	**88.23**

Table 2: Ablation study of the proposed components. CM-EMD: cross-modality earth mover's distance, CM-DL: Cross-Modality Discrimination Learning, Global: global-based feature, Local: local based features, MGF: global-based & local-based features.

Method	SYSU-MM01		RegDB	
	R1	mAP	R1	mAP
Ours(w/o CM-EMD)+KL	65.05	61.09	83.74	78.33
Ours w/o OTS	69.18	65.36	90.05	82.40
Ours	**73.39**	**68.56**	**94.37**	**88.23**

Table 3: Comparison of distribution alignment methods. KL: Kullback–Leibler divergence, w/o OTS: use cosine-similarities to replace the weights obtained by optimal transport strategy.

Evaluation

In this section, we conduct extensive experiments to investigate the effectiveness of the components of our model, *i.e.*,

Method	SYSU-MM01		RegDB	
	R1	mAP	R1	mAP
B	61.58	59.26	69.08	64.22
+ Center loss	63.15	62.36	71.65	68.12
+ Triplet loss	64.46	63.78	74.62	72.18
+ CM-DL	65.78	63.86	75.76	72.98
B + CM-EMD	71.26	66.59	92.86	84.99
+ Center loss	69.94	64.44	91.41	83.44
+ Triplet loss	71.21	66.39	91.02	81.83
+ CM-DL	**73.39**	**68.56**	**94.37**	**88.23**

Table 4: Comparison of CM-DL, center loss, and triplet loss. B: Baseline w/ MGF.

Method	SYSU-MM01		RegDB	
	R1	mAP	R1	mAP
Ours w/o ALF	72.39	67.78	91.94	84.95
Ours w/o TW	71.18	67.42	93.35	87.86
Ours	**73.39**	**68.56**	**94.37**	**88.23**

Table 5: Effect of accumulated local features (ALF) and the trainable weights (TW) of CM-DL.

MGS, CM-EMD, and CM-DL. Results are evaluated on the single-shot setting of the all-search model for SYSU-MM01 and the "Visible to Thermal" setting for RegDB.

Effectiveness of MGS. To verify the advantage of the proposed Multi-Granularity Structure (MGS), we show the results of the baseline model with the global-based feature, local-based features, and the multi-granularity features obtained by MGS, respectively. The comparisons are reported in Table 2 (#0 *vs* #2 *vs* #5). We can observe that using the local-based features can produce higher results than using the global-based features, especially on the RegDB dataset. In addition, using the multi-granularity features can further improve the results by a large margin, demonstrating the effectiveness of the proposed MGS.

Effectiveness of CM-EMD. We then evaluate the effectiveness of the proposed CM-EMD. From the results in Table 2 (#0 *vs* #1, #2 *vs* #3, #5 *vs* #6), we can find that the proposed CM-EMD can consistently improve the model performance by a large margin, no matter the types of features. For example, when using the multi-granularity features, the rank-1 accuracy of the baseline improves from 61.58% to 71.26% for SYSU-MM01, and from 69.08% to 92.86% for RegDB. These results verify the large effectiveness of our CM-EMD and show the compatibility between the proposed CM-EMD and MGS. In Table 3, we compare CM-EMD with its variant and KL divergence. For the CM-EMD variant, we directly use cosine-similarities between cross-modality sample pairs to replace the weights obtained by the optimal transport strategy. The results show that CM-EMD largely outperforms the other two methods, further demonstrating the superiority of the proposed CM-EMD. The advantage of CM-EMD over the cosine-similarity based method is mainly due to the fact that CM-EMD finds an overall optimal transport strategy to align two modalities while the cosine-similarity based method only considers the relationship between individual pairs.

Effectiveness of CM-DL. We next investigate the superiority of the proposed CM-DL. Since the loss of CM-DL is calculated on the holistic feature, which is generated based on the local-based features, we add CM-DL to models that include local-based features. The results in Table 2 (#3 *vs* #4, #6 *vs* #7) show that injecting CM-DL into the model can obtain consistent improvements. This demonstrates the effectiveness of the proposed CM-DL and also validates the compatibility of the proposed three techniques (MGS, CM-EMD, and CM-DL). In Table 4, we compare CM-DL with two popular metric learning methods (center loss (Wen et al. 2016) and triplet loss (Hermans, Beyer, and Leibe 2017)) under the 'Baseline w/ MGF' and 'Baseline w/ MGF + CM-EMD', respectively. We find that: 1) Center and triplet loss and our CM-DL can improve the results of 'Baseline w/ MGF'. 2) Center and triplet losses can not improve the results under the 'Baseline w/ MGF + CM-EMD', while our CM-DL achieves improvements on both datasets and clearly exceeds the other two methods. This further validates the benefits of our CM-DL in assisting CM-EMD. The main reason is that CM-DL is specifically designed to overcome the discrimination degradation problem caused by modality alignment, thus producing an improvement when adding to CM-EMD. However, given a model with a proper modality alignment, the constraints of center and triplet losses are mostly satisfied during training and thus will not bring further improvement without a more careful sampling strategy.

Effectiveness of accumulated local features. We evaluate the Effectiveness of accumulated local features in Table 5. We can find that accumulated local features can further improve the performance on two datasets, especially on RegDB. It is because the accumulated local features include more multi-granularity features, and improve the discrimination of representation.

Effectiveness of trainable weights. For CM-DL, the holistic feature (Eq. 10) is obtained by concatenating local features with trainable weights. We compare the results of models with and without using trainable weights. When removing the trainable weights, we use equal weights for all parts. Table 5 shows that using trainable weights can consistently produce higher results on both datasets.

Conclusion

In this paper, we propose a novel modality alignment method for visible thermal person re-identification, called Cross-Modality Earth Mover's Distance (CM-EMD), which can alleviate the impact of intra-identity variations and thus can achieve a more effective distribution alignment. Moreover, we introduce two methods to facilitate the benefit of the proposed CM-EMD, which are Cross-Modality Discrimination Learning (CM-DL) and Multi-Granularity Structure (MGS). CM-DL enables the model to learn more discriminative representation while MGS enables us to perform modality alignment in both coarse-grained to fine-grained levels. Experiments verify the advantages of the proposed methods. Our final solution can achieve state-of-the-art results on the SYSU-MM01 and RegDB datasets.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 61876159, 62076210, 62276221), the Natural Science Foundation of Fujian Province of China (No. 2022J01002), and the Science and Technology Plan Project of Xiamen (No. 3502Z20221025).

References

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein generative adversarial networks. In *ICML*.

Chen, L.; Dai, S.; Tao, C.; Shen, D.; Gan, Z.; Zhang, H.; Zhang, Y.; and Carin, L. 2018. Adversarial text generation via feature-mover's distance. In *NeurIPS*.

Chen, L.; Zhang, Y.; Zhang, R.; Tao, C.; Gan, Z.; Zhang, H.; Li, B.; Shen, D.; Chen, C.; and Carin, L. 2019a. Improving sequence-to-sequence learning via optimal transport. *arXiv preprint arXiv:1901.06283*.

Chen, X.; Wang, S.; Long, M.; and Wang, J. 2019b. Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. In *ICML*.

Chen, Y.; Wan, L.; Li, Z.; Jing, Q.; and Sun, Z. 2021. Neural Feature Search for RGB-Infrared Person Re-Identification. In *CVPR*.

Choi, S.; Lee, S.; Kim, Y.; Kim, T.; and Kim, C. 2020. Hi-CMD: Hierarchical Cross-Modality Disentanglement for Visible-Infrared Person Re-Identification. In *CVPR*.

Cuturi, M. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. In *NeurIPS*.

Dai, P.; Ji, R.; Wang, H.; Wu, Q.; and Huang, Y. 2018. Cross-Modality Person Re-Identification with Generative Adversarial Training. In *IJCAI*.

Feng, Z.; Lai, J.; and Xie, X. 2019. Learning Modality-Specific Representations for Visible-Infrared Person Re-Identification. *TIP*.

Fu, C.; Hu, Y.; Wu, X.; Shi, H.; Mei, T.; and He, R. 2021. Cm-nas: Cross-modality neural architecture search for visible-infrared person re-identification. In *ICCV*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *NeurIPS*.

Hao, X.; Zhao, S.; Ye, M.; and Shen, J. 2021. Cross-Modality Person Re-Identification via Modality Confusion and Center Aggregation. In *ICCV*.

Hao, Y.; Wang, N.; Gao, X.; Li, J.; and Wang, X. 2019a. Dual-alignment Feature Embedding for Cross-modality Person Re-identification. In *ACM MM*.

Hao, Y.; Wang, N.; Li, J.; and Gao, X. 2019b. HSME: hypersphere manifold embedding for visible thermal person re-identification. In *AAAI*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

Hermans, A.; Beyer, L.; and Leibe, B. 2017. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*.

Huang, Z.; Liu, J.; Li, L.; Zheng, K.; and Zha, Z.-J. 2022. Modality-Adaptive Mixup and Invariant Decomposition for RGB-Infrared Person Re-Identification. In *AAAI*.

Jia, M.; Zhai, Y.; Lu, S.; Ma, S.; and Zhang, J. 2020. A Similarity Inference Metric for RGB-Infrared Cross-Modality Person Re-identification. In *IJCAI*.

Li, D.; Wei, X.; Hong, X.; and Gong, Y. 2020. Infrared-Visible Cross-Modal Person Re-Identification with an X Modality. In *AAAI*.

Li, P. 2013. Tensor-sift based earth mover's distance for contour tracking. *JMIV*.

Lin, Y.; Dong, X.; Zheng, L.; Yan, Y.; and Yang, Y. 2019. A bottom-up clustering approach to unsupervised person re-identification. In *AAAI*.

Ling, Y.; Luo, Z.; Lin, Y.; and Li, S. 2021. A Multi-Constraint Similarity Learning with Adaptive Weighting for Visible-Thermal Person Re-Identification. In *IJCAI*.

Ling, Y.; Zhong, Z.; Luo, Z.; Rota, P.; Li, S.; and Sebe, N. 2020. Class-Aware Modality Mix and Center-Guided Metric Learning for Visible-Thermal Person Re-Identification. In *ACM MM*.

Liu, H.; Cheng, J.; Wang, W.; Su, Y.; and Bai, H. 2020. Enhancing the discriminative feature learning for visible-thermal cross-modality person re-identification. *Neurocomputing*.

Liu, H.; Ma, S.; Xia, D.; and Li, S. 2021. SFANet: A Spectrum-Aware Feature Augmentation Network for Visible-Infrared Person Reidentification. *TNNLS*.

Liu, H.; Tan, X.; and Zhou, X. 2020. Parameter sharing exploration and hetero-center triplet loss for visible-thermal person re-identification. *TMM*.

Lu, Y.; Wu, Y.; Liu, B.; Zhang, T.; Li, B.; Chu, Q.; and Yu, N. 2020. Cross-modality Person re-identification with Shared-Specific Feature Transfer. In *CVPR*.

Luo, H.; Jiang, W.; Gu, Y.; Liu, F.; Liao, X.; Lai, S.; and Gu, J. 2019. A strong baseline and batch normalization neck for deep person re-identification. *TMM*.

Nguyen, D.; Hong, H.; Kim, K.; and Park, K. 2017. Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors*.

Nikolentzos, G.; Meladianos, P.; and Vazirgiannis, M. 2017. Matching node embeddings for graph similarity. In *AAAI*.

Park, H.; Lee, S.; Lee, J.; and Ham, B. 2021. Learning by Aligning: Visible-Infrared Person Re-identification using Cross-Modal Correspondences. In *ICCV*.

Peyré, G.; Cuturi, M.; et al. 2019. Computational optimal transport: With applications to data science. *FTML*.

Pu, N.; Chen, W.; Liu, Y.; Bakker, E. M.; and Lew, M. S. 2020. Dual Gaussian-based Variational Subspace Disentanglement for Visible-Infrared Person Re-Identification. In *ACM MM*.

Radenović, F.; Tolias, G.; and Chum, O. 2018. Fine-tuning CNN image retrieval with no human annotation. *TPAMI*.

Rubner, Y.; Tomasi, C.; and Guibas, L. J. 1998. A metric for distributions with applications to image databases. In *ICCV*.

Rubner, Y.; Tomasi, C.; and Guibas, L. J. 2000. The earth mover's distance as a metric for image retrieval. *IJCV*.

Salimans, T.; Zhang, H.; Radford, A.; and Metaxas, D. 2018. Improving GANs using optimal transport. In *ICLR*.

Schulter, S.; Vernaza, P.; Choi, W.; and Chandraker, M. 2017. Deep network flow for multi-object tracking. In *CVPR*.

Sun, Y.; Zheng, L.; Yang, Y.; Tian, Q.; and Wang, S. 2018. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*.

Tian, X.; Zhang, Z.; Lin, S.; Qu, Y.; Xie, Y.; and Ma, L. 2021. Farewell to Mutual Information: Variational Distillation for Cross-Modal Person Re-Identification. In *CVPR*.

Wang, G.; Yuan, Y.; Chen, X.; Li, J.; and Zhou, X. 2018. Learning Discriminative Features with Multiple Granularities for Person Re-Identification. In *ACM MM*.

Wang, G.; Zhang, T.; Cheng, J.; Liu, S.; Yang, Y.; and Hou, Z. 2019a. RGB-Infrared Cross-Modality Person Re-Identification via Joint Pixel and Feature Alignment. In *ICCV*.

Wang, G.-A.; Yang, T. Z.; Cheng, J.; Chang, J.; Liang, X.; Hou, Z.; et al. 2020. Cross-Modality Paired-Images Generation for RGB-Infrared Person Re-Identification. In *AAAI*.

Wang, Z.; Wang, Z.; Zheng, Y.; Chuang, Y.-Y.; and Satoh, S. 2019b. Learning to reduce dual-level discrepancy for infrared-visible person re-identification. In *CVPR*.

Wei, Z.; Yang, X.; Wang, N.; and Gao, X. 2021a. Flexible Body Partition-Based Adversarial Learning for Visible Infrared Person Re-Identification. *TNNLS*.

Wei, Z.; Yang, X.; Wang, N.; and Gao, X. 2021b. Syncretic Modality Collaborative Learning for Visible Infrared Person Re-Identification. In *ICCV*.

Wen, Y.; Zhang, K.; Li, Z.; and Qiao, Y. 2016. A discriminative feature learning approach for deep face recognition. In *ECCV*.

Wu, A.; Zheng, W.-S.; Yu, H.-X.; Gong, S.; and Lai, J. 2017. Rgb-infrared cross-modality person re-identification. In *ICCV*.

Wu, G.; Zhu, X.; and Gong, S. 2020. Tracklet self-supervised learning for unsupervised person re-identification. In *AAAI*.

Wu, Q.; Dai, P.; Chen, J.; Lin, C.-W.; Wu, Y.; Huang, F.; Zhong, B.; and Ji, R. 2021. Discover Cross-Modality Nuances for Visible-Infrared Person Re-Identification. In *CVPR*.

Yang, F.; Wang, Z.; Xiao, J.; and Satoh, S. 2020. Mining on Heterogeneous Manifolds for Zero-Shot Cross-Modal Image Retrieval. In *AAAI*.

Yang, M.; Huang, Z.; Hu, P.; Li, T.; Lv, J.; and Peng, X. 2022. Learning with Twin Noisy Labels for Visible-Infrared Person Re-Identification. In *CVPR*.

Ye, H.; Liu, H.; Meng, F.; and Li, X. 2020a. Bi-directional exponential angular triplet loss for RGB-infrared person re-identification. *TIP*.

Ye, M.; Lan, X.; Leng, Q.; and Shen, J. 2020b. Cross-modality person re-identification via modality-aware collaborative ensemble learning. *TIP*.

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

Ye, M.; Lan, X.; Wang, Z.; and Yuen, P. C. 2019. Bi-directional Center-Constrained Top-Ranking for Visible Thermal Person Re-Identification. *TIFS*.

Ye, M.; Ruan, W.; Du, B.; and Shou, M. Z. 2021a. Channel Augmented Joint Learning for Visible-Infrared Recognition. In *ICCV*.

Ye, M.; Shen, J.; Crandall, D. J.; Shao, L.; and Luo, J. 2020c. Dynamic Dual-Attentive Aggregation Learning for Visible-Infrared Person Re-Identification. In *ECCV*.

Ye, M.; Shen, J.; Lin, G.; Xiang, T.; Shao, L.; and Hoi, S. C. 2021b. Deep learning for person re-identification: A survey and outlook. *TPAMI*.

Ye, M.; Wang, Z.; Lan, X.; and Yuen, P. C. 2018b. Visible Thermal Person Re-Identification via Dual-Constrained Top-Ranking. In *IJCAI*.

Zhang, C.; Cai, Y.; Lin, G.; and Shen, C. 2020. DeepEMD: Few-Shot Image Classification With Differentiable Earth Mover's Distance and Structured Classifiers. In *CVPR*.

Zhang, L.; Du, G.; Liu, F.; Tu, H.; and Shu, X. 2021a. Global-local multiple granularity learning for cross-modality visible-infrared person reidentification. *TNNLS*.

Zhang, Q.; Lai, C.; Liu, J.; Huang, N.; and Han, J. 2022. FMCNet: Feature-Level Modality Compensation for Visible-Infrared Person Re-Identification. In *CVPR*.

Zhang, S.; Yang, Y.; Wang, P.; Liang, G.; Zhang, X.; and Zhang, Y. 2021b. Attend to the difference: Cross-modality person re-identification via contrastive correlation. *TIP*.

Zhao, Q.; Yang, Z.; and Tao, H. 2008. Differential earth mover's distance with its applications to visual tracking. *TPAMI*.

Zheng, K.; Lan, C.; Zeng, W.; Zhang, Z.; and Zha, Z.-J. 2021. Exploiting sample uncertainty for domain adaptive person re-identification. In *AAAI*.