

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

Yongguo Ling, Zhun Zhong, Donglin Cao, Zhiming Luo, Yaojin Lin, Shaozi Li, Nicu Sebe

Abstract—Visible thermal person re-identification (VT-ReID) suffers from the inter-modality discrepancy and intra-identity variations. **Distribution alignment** is a popular solution for VT-ReID, which, however, 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, the **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.

Index Terms—Visible-Thermal Person Re-Identification, Cross-Modality, Earth Mover's Distance, Modality Alignment.

I. 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 [1], [2] 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) [3]. 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).

Distribution alignment [4]–[6] 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

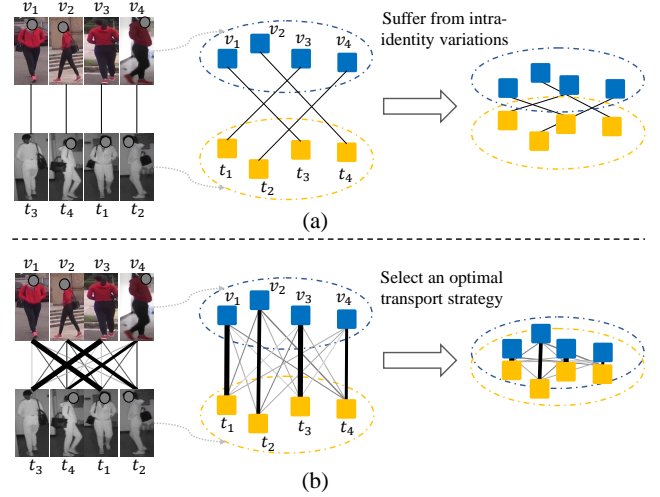


Fig. 1: (a) Traditional distribution alignment. The randomly selected pairs usually have high intra-identity variations, hindering the reduction of inter-modality discrepancy. (b) Distribution alignment with Cross-Modality Earth Mover's Distance (CM-EMD). Our CM-EMD assigns large weights to pairs that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

a training mini-batch, the randomly selected cross-modality pairs of the same identity commonly have a large variation in pose, view and background (Figure 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, the 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 [7] to select an optimal transport strategy across two modalities, which assigns a large transport weight between two cross-modality samples that have a smaller intra-identity variation (Figure 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. On the one hand, since CM-EMD mainly focuses on aligning the cross-modality distribution, this will inevitably degrade the discrimination ability of the representation. To overcome this issue, we present a **Cross-Modality Discrimination Learning** (CM-DL), which can improve the discrimination by reducing the ratio between the intra-identity and the inter-identity variances. In addition, knowing that local information is important for Re-ID [8], we further introduce a **Multi-Granularity Structure** (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 on 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), versatily 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.

II. RELATED WORK

Visible-Thermal person re-identification (VT-ReID) aims at matching a thermal / visible query person from a visible / thermal gallery. Generally, existing methods of VT-ReID can be mainly divided into four groups, *i.e.*, feature extraction based methods, metric learning based methods, distribution alignment based methods, and image generation based methods.

Feature Extractor based methods mainly focus on designing a cross-modality network to extract modality-invariant and discriminative representation [9]–[17]. Wu et al. [9] first consider the visible-thermal problem in ReID and introduce a visible-thermal dataset (SYSU-MM01). In addition, they inject a deep zero-padding approach into the network for evolving domain-specific nodes. Ye et al. [10] introduce a two-stream network to handle the two modalities respectively and a dual-constrained top-ranking loss for learning robust cross-modality embedding. However, Yang et al. [12] point out that the training of a two-stream model may distort the distribution of unseen classes. To tackle this issue, a bi-directional random walk scheme is proposed to discover more reliable cross-modality feature distributions. Jia et al. [15] design a similarity inference metric (SIM) to leverage the similarities between intra-modality samples to help reduce cross-modality gap. Lu et al. [13] introduce a cross-modality feature transfer algorithm which can effectively extract the modality-shared information

and modality-specific characteristic in the feature space. To extract more fine-grained features, Ye et al. [14] develop a dynamic dual-attentive aggregation module by mining both intra-modality part-level and cross-modality graph-level contextual. Recently, Chen et al. [16] propose Neural Feature Search (NFS) to automate select identity-related feature for matching. Tian et al. [17] introduce Variational Self-Distillation (VSD) to fitting the mutual information and reducing view and irrelevant information, learning more robustness representation. These approaches design different feature extractors to extract either global-level feature or part-level feature. However, they ignore jointly considered these two aspects and studied their mutual benefits. Therefore, we devise a framework to extract and combine global- and part-level features, which can carry more discriminative content.

Metric learning based methods are proposed to learn an embedding space by explicitly enforcing the intra-class samples of two modalities close to each other [18]–[21]. Ye et al. [18] introduce a hierarchical inter-modality metric learning to learn cross-modality embedding. Hao et al. [19] map the features to a hyper-sphere manifold which can effectively reduce the intra-modality variations and cross-modality variations with joint classification and identification constraints. In addition, Ling et al. [21] introduce a center-guided loss function by formulating learning constraints among class centers and instances, which can effectively reduce inter- and intra- modalities discrepancy. Different from these approaches, this paper exploit the Cross-modality Earth Mover's Distance (cmEMD) to reduce cross-modality discrepancy.

Distribution alignment based methods are mainly proposed to decrease the distribution divergence to smooth the inter-modality discrepancy and gain modality invariant feature [4]–[6], [22]. Dai et al. [6] introduce a generative adversarial network to design a modality discriminator to learn discriminative feature. Hao et al. [5] propose dual-alignment to reduce the modality gap in the concept of spatial and modality by extracting part-level features and learning with distribution and correlation losses. Pu et al. [4] propose a dual Gaussian-based variational auto-encoder (DG-VAE), which allows model to explore unobserved data enforce both modalities to follow the true distribution. In order to explore nuances information, Wu et al. [22] introduce MPANet to align modality and extract discriminative features by a modality alleviation module and a pattern alignment module. These approaches can alleviate the large inter-modality discrepancy. Nevertheless, the intra-modality variations may limit the distribution alignment of two modalities. Therefore, we propose the Cross-Modality Earth Mover's Distance (cmEMD) to measure distance of two graphs of each modality and align these two modalities.

Image generation based methods commonly attempt to generate fake/virtual images that are used to bridge the modality gap in the image-level [23]–[27]. Wang et al. [23] design a modality translation network to transfer the styles of images from one modality to another one. In addition, they unify the image and its transferred counterpart into a multi-spectral image to further reduce the appearance discrepancy during representation learning. Similarly, Wang et al. [24] propose the Alignment Generative Adversarial Network (AlignGAN) to

decrease intra-class variations in both pixel-level and feature-level. Instead of transferring images between two modalities, Li et al. [25] produce an auxiliary X modality and convert the cross-modality learning as an X-Infrared-Visible three-mode learning problem. To reduce the impact of id-unrelated factors, several works ([26], [27]) utilize the variational autoencoder and generative adversarial network to disentangle feature into id-related feature and id-unrelated feature, where the id-related feature is used for robust cross-modality matching. Despite the effectiveness of the above methods, they commonly require to train generative adversarial networks, which are difficult to optimize and sensitive to parameter settings.

Earth Mover's Distance (EMD) [7] 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 [28], tracking [29]–[31], graph matching [32], document retrieval [33], overcoming mode collapsing [34]–[36] of Generative Adversarial Network (GAN) [37], and few-shot learning [38]. Different from them, this work is the first of employing EMD to solve the problem of VT-ReID, where we propose the cross-modality EMD to effectively learn modality-invariant representation.

III. METHODOLOGY

Overview. The framework of the proposed method is shown in Fig. 2. We first extract global feature, 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 compute 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.

A. 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.

B. 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 less 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 caused by various factors while we only have the identity information. On the other hand, EMD [7] 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 the 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 less 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. III-D. The feature distributions of visible modality and thermal modality are denoted as $\nu \in \mathcal{P}(\mathbb{F}^v)$ and $\tau \in \mathcal{P}(\mathbb{F}^t)$, respectively. $\Pi(\nu, \tau)$ represents all joint distributions $\rho(\mathbf{f}^v, \mathbf{f}^t)$, where $\mathbf{f}^v \in \mathbb{F}^v$ and $\mathbf{f}^t \in \mathbb{F}^t$. The CM-EMD among the samples of two modalities can be defined as:

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

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

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

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

$$\begin{aligned} S_{ij} \geq 0, \quad \sum_{j=1}^{N^t} S_{ij} &= v_i, \quad \sum_{i=1}^{N^v} S_{ij} = t_j, \\ \forall i &= 1, \dots, N^v, \quad \forall j = 1, \dots, N^t. \end{aligned} \quad (3)$$

To the reduce computational complexity, we use Sinkhorn algorithm [40], [41] to solve the Eq. 2.

Discussion. After optimization, there is a negative correlation between the transport weight S_{ij} and the transport cost $M(\mathbf{f}_i^v, \mathbf{f}_j^t)$. That is, the two samples, \mathbf{f}_i^v and \mathbf{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

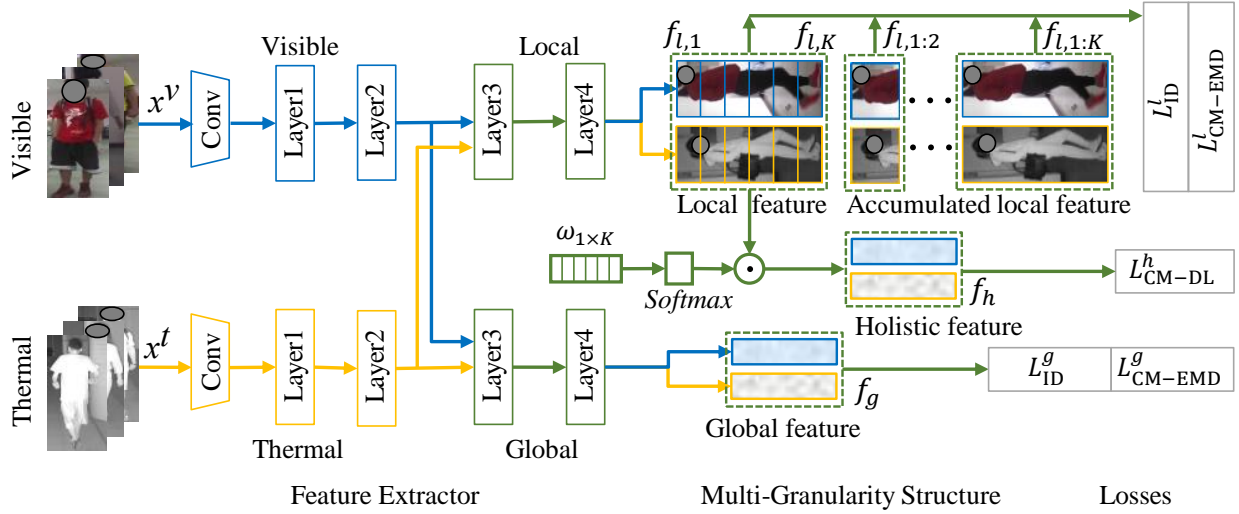


Fig. 2: The framework of the proposed method. In the shallow layers, we use different parameters for visible modality and thermal modality. The shallow layers are the first convolutional layer, 1-th residual block and 2-th residual block of the ResNet-50 [39]. After the shallow layers, we design two 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 [39], which are shared by both modalities. The parameters of the two streams are different. Given the inputs, we first extract global feature, local features and accumulated local features with our Multi-Granularity Structure (MGS). We then calculate the losses of identity classification (L_{ID}^g and L_{ID}^l) and the losses of our Cross-Modality Earth Mover's Distance (L_{CM-EMD}^g and L_{CM-EMD}^l) 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 our Cross-Modality Discrimination Learning (L_{CM-DL}^h) is calculated with the holistic feature.

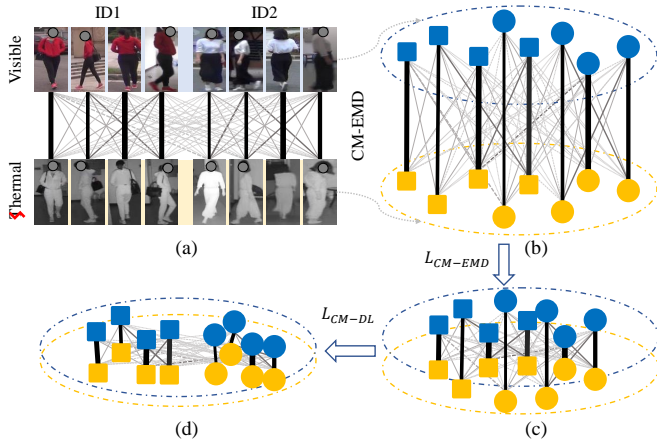


Fig. 3: Illustration of the proposed CM-EMD (a→c) and CM-DL (c→d). CM-EMD can effectively reduce the modality gap while CM-DL can effectively promote the discrimination ability of the model. The pairs with high weights are connected with wider lines. Shapes indicate the identities. Colors represent the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

on reducing the discrepancy of samples that have less 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).

C. Cross-Modality Discrimination Learning

One disadvantage of modality alignment is that the optimization process may degrade the discrimination capability of the model, which is also a problem of the proposed CM-EMD. 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:

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

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

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

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:

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

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

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

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

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

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).

D. Multi-Granularity Structure

In the community of ReID [8], it has been demonstrated that local information is critical for extracting discriminative pedestrian representation. Inspired by this, we propose the multi-granularity structure (MGS) to obtain more discriminative feature as well as align the modalities in both coarse- and fine-grained aspects.

Given the feature of the last residual convolution layer $f_{res} \in \mathbb{R}^{H,W,C}$, we extract three types of features: global feature, local feature, accumulated local feature, which are introduced in the followings. Note that, since we adopt different streams to obtain f_{res} for global feature and local-based features, we use $f_{g,res}$ for global stream and $f_{l,res}$ for local stream.

Global Feature. We directly apply the generalized-mean pooling (GeM) [42] and the batch normalization neck (BN-Neck) [43] on $f_{g,res}$ and obtain the global feature $f_g \in \mathbb{R}^{1,C}$.

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, $\{f_{l,1}, \dots, f_{l,K}\}$.

Accumulated Local Feature. To obtain more diverse features, we extract accumulated local features by combining different number of part features, which are represented by $\{f_{l,1:2}, \dots, f_{l,1:K}\}$. 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:

$$L_{CM-EMD}^g = D_{CM-EMD}(\mathbb{F}_g^v, \mathbb{F}_g^t), \\ L_{CM-EMD}^l = \sum_{k=1}^K D_{CM-EMD}(\mathbb{F}_{l,k}^v, \mathbb{F}_{l,k}^t) + \alpha \sum_{k=2}^K D_{CM-EMD}(\mathbb{F}_{l,1:k}^v, \mathbb{F}_{l,1:k}^t), \quad (9)$$

where α is the hyper-parameter and K is the part number. L_{CM-EMD}^g is the global-based loss of CM-EMD and L_{CM-EMD}^l is the local-based loss of CM-EMD.

Loss for 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 to the calculated loss of CM-DL. Specifically, the holistic feature is represented as:

$$f_h = [\omega_1 f_{l,1} \mid \omega_2 f_{l,2} \mid \dots \mid \omega_K f_{l,K}], \quad (10)$$

where $[\cdot \mid \cdot]$ indicates the concatenation operation. $\omega_1, \dots, \omega_K$ are the trainable weights, which are normalized by the Soft-Max function.

The loss of CM-DL is re-formulated as follows:

$$L_{CM-DL}^h = \frac{V_{intra}(\mathbb{F}_h^v, \mathbb{F}_h^t)}{V_{inter}(\mathbb{F}_h^v, \mathbb{F}_h^t)}. \quad (11)$$

Losses for Identity Classification. As explained in Sec. III-A, 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 feature, K for local features, and K for accumulated local features), which are fully-connected layers with output length of #identities, after these three types of features. We then calculated identity classification losses according to the outputs of these classification heads, formulated as:

$$L_{ID}^g = \text{CrossEntropy}(\mathbb{H}_g, \mathbb{Y}), \\ L_{ID}^l = \sum_{k=1}^K \text{CrossEntropy}(\mathbb{H}_{l,k}, \mathbb{Y}) + \alpha \sum_{k=2}^K \text{CrossEntropy}(\mathbb{H}_{l,1:k}, \mathbb{Y}), \quad (12)$$

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

E. Overall

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

$$\arg \min_{\theta} \gamma_1 L_{CM-DL}^h + \gamma_2 L_{ID}^l + \gamma_3 L_{CM-EMD}^l + \gamma_4 L_{ID}^g + \gamma_5 L_{CM-EMD}^g, \quad (13)$$

where $\gamma_{1 \rightarrow 5}$ 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:

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

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

IV. EXPERIMENT

A. Experimental settings

Datasets. Experiments are conducted on two VT-ReID datasets, i.e., SYSU-M001 [9] and RegDB [52].

SYSU-M001 [9] contains 287,628 RGB images and 15,729 infrared images, which are captured by four RGB cameras and

TABLE I: Comparison with the state-of-the-art methods on the RegDB dataset.

Settings Method	Visible to Thermal				Thermal to Visible			
	R1	R10	R20	mAP	R1	R10	R20	mAP
HOG [44]	13.49	33.22	43.66	10.31	/	/	/	/
LOMO [45]	0.85	2.47	4.1	2.28	/	/	/	/
One-stream [9]	13.11	32.98	42.51	14.02	/	/	/	/
Two-stream [9]	12.43	30.36	40.96	13.42	/	/	/	/
Zero-Padding [9]	17.75	34.21	44.35	18.9	/	/	/	/
HCML [18]	24.44	47.53	56.78	20.8	21.70	45.02	55.58	22.24
BDTR [10]	33.47	58.42	67.52	31.83	32.92	58.46	68.43	31.96
DGD+MSR [20]	48.43	70.32	79.95	48.67	/	/	/	/
D2RL [23]	/	/	/	/	43.4	66.1	76.3	44.1
EDFL [11]	52.58	72.1	81.47	52.98	51.89	72.09	81.04	52.13
D-HSME [19]	50.85	73.36	81.66	47	50.15	72.40	81.07	46.16
AlignGAN [24]	57.9	/	/	53.6	56.3	/	/	53.4
DFE [5]	70.13	86.32	91.96	69.14	67.99	85.56	91.41	66.70
Hi-CMD [26]	/	/	/	/	70.93	86.39	/	66.04
PIG [27]	48.50	/	/	49.3	48.1	/	/	48.90
Xmodal [25]	62.21	83.13	91.72	60.18	/	/	/	/
CMM-CML [21]	59.81	80.39	88.69	60.86	/	/	/	/
DDAG [14]	69.34	86.19	91.49	63.46	68.06	85.15	90.31	61.80
DG-VAE [4]	72.97	/	86.89	71.78	/	/	/	/
cm-SSFT [13]	73.3	/	/	72.9	71.0	/	/	71.7
SIM [15]	75.29	/	/	74.47	78.30	/	/	75.24
AGW [46]	70.05	86.21	91.55	66.37	70.49	87.12	91.84	65.90
CMAAlign [47]	74.17	/	/	67.64	72.43	/	/	65.46
SFANet [48]	76.31	91.02	94.27	68.00	70.15	85.24	89.27	63.77
NFS [16]	80.54	91.96	95.07	72.1	77.95	90.45	93.62	69.79
VCD-VML [17]	73.2	/	/	71.6	71.8	/	/	70.1
MCLNet [49]	80.3	92.7	96.03	73.07	75.93	90.93	94.59	69.49
MPANet [22]	83.7	/	/	80.9	82.8	/	/	80.7
SMCL [50]	83.93	/	/	79.83	83.05	/	/	78.57
CM-NAS [51]	84.54	95.18	97.85	80.32	82.57	94.51	97.37	78.31
Ours	94.37	98.93	99.42	88.23	92.77	98.50	99.66	86.85

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 setting for these two modes.

RegDB [52] 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), rank-10 (R10), and rank-20 (R20) accuracies.

Implementation Details. The baseline model is trained with only the identity classification loss. 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.*, $N_c^v = N_c^t = 8$) for SYSU-M001, and 4 RGB images and 4 infrared images (*i.e.*, The input images are resized to $384 \times 192 \times 3$. We choose the ResNet-50 [39] as the backbone of the feature extractor. The SGD optimizer is used to update the parameters of the network, where we set initial learning rate to 0.01 and divide the learning rate by 10 after every 30 epochs. We train the model for a total of 80 epochs. For the hyper-parameters of the proposed model,

we set α (Eq. 9) to 0.2 and 1.0 for SYSU-M001 and RegDB, respectively. $\gamma_{1 \rightarrow 5}$ (Eq. 13) are to $\{1, 1, 0.1, 2, 0.1\}$ for SYSU-M001 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-M001 and RegDB, respectively.

B. 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. The competitors include feature extraction based methods (Zero-Padding [9], BDTR [10], cm-SSFT [13], SIM [15], DDAG [14], NFS [16], VCD-VML [17]), metric learning based methods (HCML [18], D-HSME [19], BFE+HPI [53], DGD+MSR [20], CMM-CML [21], MCLNet [49]), distribution alignment based methods (cmGAN [6], DFE [5], DG-VAE [4], MPANet [22]), and image generation based methods (D2RL [23], AlignGAN [24], Xmodal [25], Hi-CMD [26], PIG [27]).

Results on RegDB. We first compare our method with state-of-the-art methods on RegDB. The results are reported in Tabel I. It is clear that our method outperforms the state-of-the-art methods by a large margin on both evaluation settings. Specifically, we obtain **rank-1 accuracy = 94.37%** and **mAP accuracy = 88.23%** for the “Visible to Thermal” setting, and, **rank-1 accuracy = 92.77%** and **mAP accuracy = 86.85%** for “Thermal to Visible” setting. Compared to the current best competitor (CM-NAS [51] and SMCL [50] published in ICCV 2021), we show significant improvements. For example, our method is higher than CM-NAS by 9.83% in rank-1 accuracy and 7.91% in mAP accuracy on the “Visible to Thermal”

TABLE II: Comparison with the state-of-the-art methods on the SYSU-MM01 dataset.

Settings	All-search				Indoor-search			
Method	R1	R10	R20	mAP	R1	R10	R20	mAP
HOG [44]	2.76	18.25	31.91	4.24	3.22	24.68	44.52	7.25
LOMO [45]	3.64	23.18	37.28	4.53	5.75	34.35	54.9	10.19
One-stream [9]	12.04	49.68	66.74	13.67	16.94	63.55	82.1	22.95
Two-stream [9]	11.65	47.99	65.5	12.85	15.6	61.18	81.02	21.49
Zero-padding [9]	14.8	54.12	71.33	15.95	20.58	68.38	85.79	26.92
HCML [18]	14.32	53.16	69.17	16.16	/	/	/	/
cmGAN [6]	26.97	67.51	80.56	27.8	31.63	77.23	89.18	42.19
BDTR [10]	17.01	55.43	71.96	19.66	/	/	/	/
D-HSME [19]	20.68	62.74	77.95	23.12	/	/	/	/
D2RL [23]	28.9	70.6	82.4	29.2	/	/	/	/
DGD+MSR [20]	37.35	83.4	93.44	38.11	39.64	89.29	97.66	50.88
BFE+HPI [53]	41.36	84.78	94.51	42.95	45.77	91.82	98.46	56.52
AlignGAN [24]	42.4	85	93.7	40.7	45.9	87.6	94.4	54.3
DFE [5]	48.71	88.86	95.27	48.59	52.25	89.86	95.85	59.68
EDFL [11]	36.94	84.52	93.22	40.77	/	/	/	/
Hi-CMD [26]	34.94	77.58	/	35.94	/	/	/	/
PIG [27]	38.1	80.7	89.9	36.9	43.8	86.2	94.2	52.9
Xmodal [25]	49.92	89.79	95.96	50.73	/	/	/	/
CMM-CML [21]	51.8	92.72	97.71	51.21	54.98	94.38	99.41	63.7
DDAG [14]	54.75	90.39	95.81	53.02	61.02	94.06	98.41	67.98
DG-VAE [4]	59.49	/	93.77	58.46	/	/	/	/
SIM [15]	60.88	/	/	56.93	/	/	/	/
cm-SSFT [13]	61.6	89.2	93.9	63.2	70.5	94.9	97.7	72.6
AGW [46]	47.50	84.39	92.14	47.65	54.17	91.14	95.98	62.97
CMAAlign [47]	55.41	/	/	54.14	58.46	/	/	66.33
NFS [16]	56.91	91.34	96.52	55.45	62.79	96.53	99.07	69.79
VCD-VML [17]	60.02	94.18	98.14	58.8	66.05	96.59	99.38	72.98
CM-NAS [51]	61.99	92.87	97.25	60.02	67.01	97.02	99.32	72.95
MC-AWL [54]	64.82	/	/	60.81	/	/	/	/
SFANet [48]	65.74	92.98	97.05	60.83	71.60	96.60	99.45	80.05
MCLNet [49]	65.4	93.33	97.14	61.98	72.56	96.98	99.2	76.58
SMCL [50]	67.39	92.87	96.76	61.78	68.84	96.55	98.77	75.56
MPANet [22]	70.58	96.21	98.8	68.24	76.74	98.21	99.57	80.95
Ours	73.39	96.24	98.82	68.56	80.53	98.31	99.91	82.71

setting, and, outperforms SMCL by 9.72% in rank-1 accuracy and 8.28% in mAP accuracy on the “Thermal to Visible” setting, respectively. This indicates that our method leads a new state-of-the-art performance on RegDB.

Results on SYSU-MM001. The comparisons on SYSU-MM001 are reported in Tabel II. In all settings, our method achieves the best results in rank-1 accuracy and obtains competitive results in mAP. Specifically, we achieve **rank-1 accuracy = 73.39%** and **mAP accuracy = 68.56%** for the all-search mode, and, **rank-1 accuracy = 80.53%** and **mAP accuracy = 82.71%** for the indoor-search mode. Compared to the current best competitor (MPANet [22] 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.

C. Evaluation

In this section, we conduct extensive experiments to investigate the effectiveness of the components of our model, *i.e.*, MGS, CM-EMD, and CM-DL. Results are evaluated on the single-shot setting of 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 III

TABLE III: 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-M001		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 IV: Effect of accumulated local features (ALF).

Method	RegDB		SYSU-M001	
	R1	mAP	R1	mAP
Ours w/o ALF	91.94	84.95	72.39	67.78
Ours	94.37	88.23	73.39	68.56

(#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 accumulated local features. We evaluate the Effectiveness of accumulated local features in Table IV. We

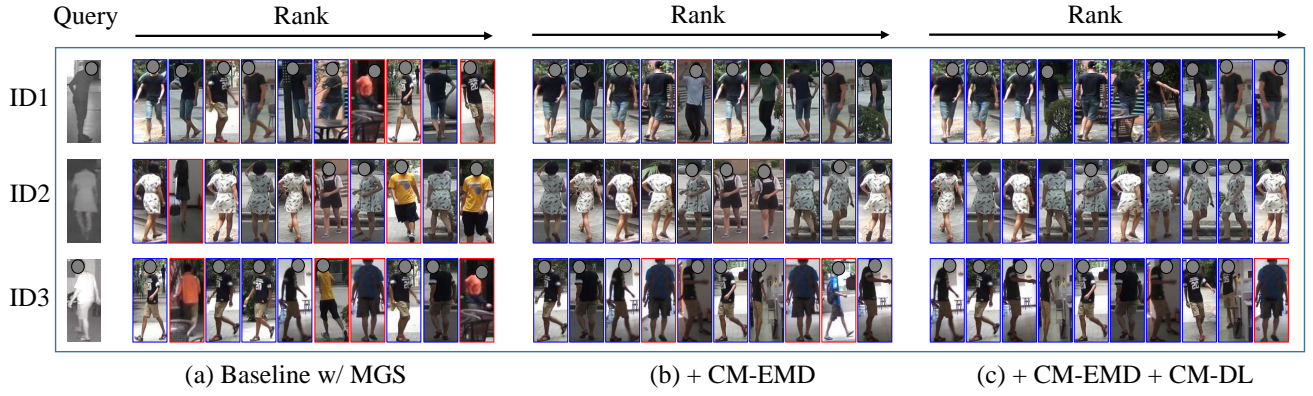


Fig. 4: Comparison of the top-10 retrieval results of the (a) Baseline w/ MGS, (b) Baseline w/ MGS + CM-EMD and (c) Baseline w/ MGS + CM-EMD + CM-DL. We evaluate the models on the SYSU-M001 dataset. Sample with blue / red box is positive / negative (best view in zoom).

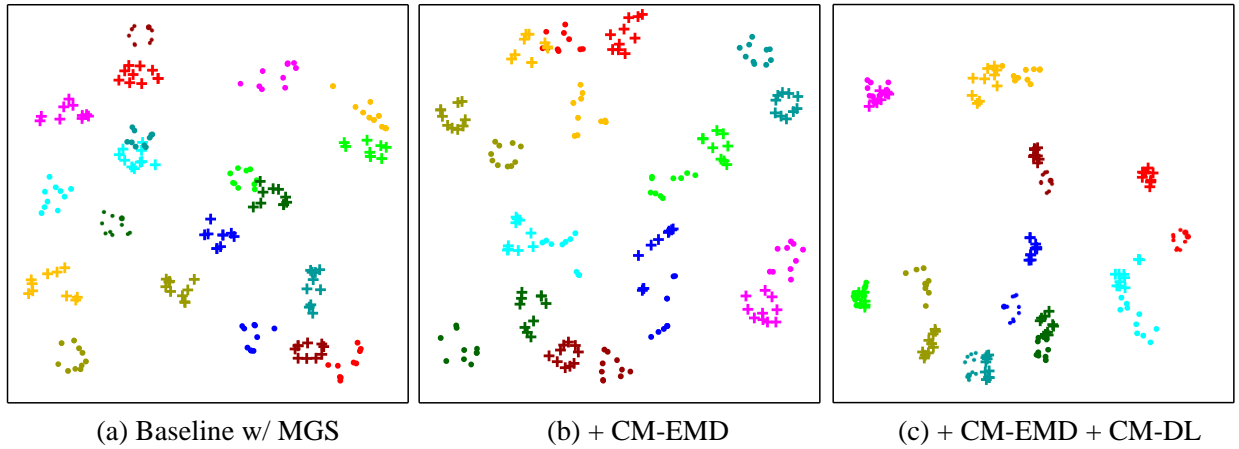


Fig. 5: The t-SNE [55] of learned features for RegDB of (a) Baseline w/ MGS, (b) + CM-EMD, and (c) + CM-EMD + CM-DL. Colors represent the identities. Circle-dot denotes the visible modality and cross-mark represents the thermal modality. For better visualization, we randomly select 10 person identities with 10 samples for each modality from the testing set.

can find that accumulated local features can further improve the performance on two datasets, especially on RegDB.

Effectiveness of CM-EMD. We then evaluate the effectiveness of the proposed CM-EMD. From the results in Table III (#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% and 69.08% to 92.86% for SYSU-M001 and RegDB, respectively. These results verify the large effectiveness of our CM-EMD and show the compatibility between the proposed CM-EMD and MGS. In Table V, we compare CM-EMD with its variant and KL divergence. For the CM-EMD variant, we directly use cosine-similarities between samples 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 cosine-similarity based method is mainly due to that CM-EMD finds an overall optimal transport strategy to align two modalities while cosine-similarity based

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

Method	SYSU-M001		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 VI: Comparison of CM-DL, center loss and triplet loss. Base: Baseline w/ MGF + CM-EMD

Method	SYSU-M001		RegDB	
	R1	mAP	R1	mAP
Base	71.26	66.59	92.86	84.99
+ Center loss [56]	69.94	64.44	91.41	83.44
+ Triplet loss [57]	71.21	66.39	91.02	81.83
+ CM-DL	73.39	68.56	94.37	88.23

method only considers the relationship between individual pairs.

Effectiveness of CM-DL. We next investigate the superi-

TABLE VII: Analysis of the trainable weights (TW) of CM-DL.

Method	SYSU-M001		RegDB	
	R1	mAP	R1	mAP
Ours w/o TW	71.18	67.42	93.35	87.86
Ours	73.39	68.56	94.37	88.23

ority 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 III (#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 method and also validates the compatibility of the proposed three techniques (MGS, CM-EMD and CM-DL). In Table VI, we compare CM-DL with two popular metric learning methods, *i.e.*, center loss [56] and triplet loss [57]. We find that center and triplet losses can not improve the results but obtain a performance decrease on both SYSU-M001 and RegDB. However, 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. It is because 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 Trainable Weights. For CM-DL, the holistic feature (Eq. 10) is obtained by concatenating local features with trainable weights. In Table VII, we report the results of models with and without using trainable weights. When removing the trainable weights, we use equal weights for all parts. Table VII shows that using trainable weights can consistently produce higher results on both datasets.

D. Visualization

To better reflect the superiority of the proposed methods, we visualize the top-10 retrieval results on the SYSU-M001 dataset. The compared methods are “Baseline w/ MGS”, “Baseline w/ MGS + CM-EMD” and “Baseline w/ MGS + CM-EMD + CM-DL”. The comparisons are illustrated in Fig. 4. It is clear that the proposed CM-EMD and CM-DL can significantly improve the ranking lists. That is, more positive samples are ranked in the top positions when using CM-EMD and CM-DL.

t-SNE of Feature Distribution. We then show the t-SNE [55] of features on the testing set for RegDB in Figure 5. Results on the baseline and our proposed CM-EMD show that CM-EMD can effectively align the cross-modality gap. The CM-DL can further increase the similarity of the inter- and intra-modality features of the same identity.

V. 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 the state-of-the-art results on SYSU-M001 and RegDB.

REFERENCES

- [1] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, “Camera style adaptation for person re-identification,” in *CVPR*, 2018. 1
- [2] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, “Person re-identification by multi-channel parts-based cnn with improved triplet loss function,” in *CVPR*, 2016. 1
- [3] M. Ye, X. Lan, Z. Wang, and P. C. Yuen, “Bi-directional center-constrained top-ranking for visible thermal person re-identification,” *TIFS*, 2019. 1
- [4] N. Pu, W. Chen, Y. Liu, E. M. Bakker, and M. S. Lew, “Dual gaussian-based variational subspace disentanglement for visible-infrared person re-identification,” in *ACM MM*, 2020. 1, 2, 6, 7
- [5] Y. Hao, N. Wang, X. Gao, J. Li, and X. Wang, “Dual-alignment feature embedding for cross-modality person re-identification,” in *ACM MM*, 2019. 1, 2, 6, 7
- [6] P. Dai, R. Ji, H. Wang, Q. Wu, and Y. Huang, “Cross-modality person re-identification with generative adversarial training,” in *IJCAI*, 2018. 1, 2, 6, 7
- [7] Y. Rubner, C. Tomasi, and L. J. Guibas, “A metric for distributions with applications to image databases,” in *ICCV*, 1998. 1, 3
- [8] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, “Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline),” in *ECCV*, 2018. 2, 5
- [9] A. Wu, W.-S. Zheng, H.-X. Yu, S. Gong, and J. Lai, “Rgb-infrared cross-modality person re-identification,” in *ICCV*, 2017. 2, 5, 6, 7
- [10] M. Ye, Z. Wang, X. Lan, and P. C. Yuen, “Visible thermal person re-identification via dual-constrained top-ranking,” in *IJCAI*, 2018. 2, 6, 7
- [11] H. Liu, J. Cheng, W. Wang, Y. Su, and H. Bai, “Enhancing the discriminative feature learning for visible-thermal cross-modality person re-identification,” *Neurocomputing*, 2020. 2, 6, 7
- [12] F. Yang, Z. Wang, J. Xiao, and S. Satoh, “Mining on heterogeneous manifolds for zero-shot cross-modal image retrieval,” in *AAAI*, 2020. 2
- [13] Y. Lu, Y. Wu, B. Liu, T. Zhang, B. Li, Q. Chu, and N. Yu, “Cross-modality person re-identification with shared-specific feature transfer,” in *CVPR*, 2020. 2, 6, 7
- [14] M. Ye, J. Shen, D. J. Crandall, L. Shao, and J. Luo, “Dynamic dual-attentive aggregation learning for visible-infrared person re-identification,” in *ECCV*, 2020. 2, 6, 7
- [15] M. Jia, Y. Zhai, S. Lu, S. Ma, and J. Zhang, “A similarity inference metric for rgb-infrared cross-modality person re-identification,” in *IJCAI*, 2020. 2, 6, 7
- [16] Y. Chen, L. Wan, Z. Li, Q. Jing, and Z. Sun, “Neural feature search for rgb-infrared person re-identification,” in *CVPR*, 2021. 2, 6, 7
- [17] X. Tian, Z. Zhang, S. Lin, Y. Qu, Y. Xie, and L. Ma, “Farewell to mutual information: Variational distillation for cross-modal person re-identification,” in *CVPR*, 2021. 2, 6, 7
- [18] M. Ye, X. Lan, J. Li, and P. C. Yuen, “Hierarchical discriminative learning for visible thermal person re-identification,” in *AAAI*, 2018. 2, 6, 7
- [19] Y. Hao, N. Wang, J. Li, and X. Gao, “Hsme: hypersphere manifold embedding for visible thermal person re-identification,” in *AAAI*, 2019. 2, 6, 7
- [20] Z. Feng, J. Lai, and X. Xie, “Learning modality-specific representations for visible-infrared person re-identification,” *TIP*, 2019. 2, 6, 7

- [21] Y. Ling, Z. Zhong, Z. Luo, P. Rota, S. Li, and N. Sebe, "Class-aware modality mix and center-guided metric learning for visible-thermal person re-identification," in *ACM MM*, 2020. 2, 6, 7
- [22] Q. Wu, P. Dai, J. Chen, C.-W. Lin, Y. Wu, F. Huang, B. Zhong, and R. Ji, "Discover cross-modality nuances for visible-infrared person re-identification," in *CVPR*, 2021. 2, 6, 7
- [23] Z. Wang, Z. Wang, Y. Zheng, Y.-Y. Chuang, and S. Satoh, "Learning to reduce dual-level discrepancy for infrared-visible person re-identification," in *CVPR*, 2019. 2, 6, 7
- [24] G. Wang, T. Zhang, J. Cheng, S. Liu, Y. Yang, and Z. Hou, "Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment," in *ICCV*, 2019. 2, 6, 7
- [25] D. Li, X. Wei, X. Hong, and Y. Gong, "Infrared-visible cross-modal person re-identification with an x modality," in *AAAI*, 2020. 2, 3, 6, 7
- [26] S. Choi, S. Lee, Y. Kim, T. Kim, and C. Kim, "Hi-cmd: Hierarchical cross-modality disentanglement for visible-infrared person re-identification," in *CVPR*, 2020. 2, 3, 6, 7
- [27] G.-A. Wang, T. Z. Yang, J. Cheng, J. Chang, X. Liang, Z. Hou *et al.*, "Cross-modality paired-images generation for rgb-infrared person re-identification," in *AAAI*, 2020. 2, 3, 6, 7
- [28] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *IJCV*, 2000. 3
- [29] Q. Zhao, Z. Yang, and H. Tao, "Differential earth mover's distance with its applications to visual tracking," *TPAMI*, 2008. 3
- [30] P. Li, "Tensor-sift based earth mover's distance for contour tracking," *JMIV*, 2013. 3
- [31] S. Schuster, P. Vernaza, W. Choi, and M. Chandraker, "Deep network flow for multi-object tracking," in *CVPR*, 2017. 3
- [32] G. Nikolentzos, P. Meladianos, and M. Vazirgiannis, "Matching node embeddings for graph similarity," in *AAAI*, 2017. 3
- [33] L. Chen, Y. Zhang, R. Zhang, C. Tao, Z. Gan, H. Zhang, B. Li, D. Shen, C. Chen, and L. Carin, "Improving sequence-to-sequence learning via optimal transport," *arXiv preprint arXiv:1901.06283*, 2019. 3
- [34] L. Chen, S. Dai, C. Tao, D. Shen, Z. Gan, H. Zhang, Y. Zhang, and L. Carin, "Adversarial text generation via feature-mover's distance," in *NeurIPS*, 2018. 3
- [35] T. Salimans, H. Zhang, A. Radford, and D. Metaxas, "Improving gans using optimal transport," in *ICLR*, 2018. 3
- [36] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *ICML*, 2017. 3
- [37] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NeurIPS*, 2014. 3
- [38] C. Zhang, Y. Cai, G. Lin, and C. Shen, "Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers," in *CVPR*, 2020. 3
- [39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016. 4, 6
- [40] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," in *NeurIPS*, 2013. 3
- [41] G. Peyré, M. Cuturi *et al.*, "Computational optimal transport: With applications to data science," *FTML*, 2019. 3
- [42] F. Radenović, G. Tolias, and O. Chum, "Fine-tuning cnn image retrieval with no human annotation," *TPAMI*, 2018. 5
- [43] H. Luo, W. Jiang, Y. Gu, F. Liu, X. Liao, S. Lai, and J. Gu, "A strong baseline and batch normalization neck for deep person re-identification," *TMM*, 2019. 5
- [44] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005. 6, 7
- [45] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *CVPR*, 2015. 6, 7
- [46] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. Hoi, "Deep learning for person re-identification: A survey and outlook," *TPAMI*, 2021. 6, 7
- [47] H. Park, S. Lee, J. Lee, and B. Ham, "Learning by aligning: Visible-infrared person re-identification using cross-modal correspondences," in *ICCV*, 2021. 6, 7
- [48] H. Liu, S. Ma, D. Xia, and S. Li, "Sfanet: A spectrum-aware feature augmentation network for visible-infrared person reidentification," *TNNLS*, 2021. 6, 7
- [49] X. Hao, S. Zhao, M. Ye, and J. Shen, "Cross-modality person re-identification via modality confusion and center aggregation," in *ICCV*, 2021. 6, 7
- [50] Z. Wei, X. Yang, N. Wang, and X. Gao, "Syncretic modality collaborative learning for visible infrared person re-identification," in *ICCV*, 2021. 6, 7
- [51] C. Fu, Y. Hu, X. Wu, H. Shi, T. Mei, and R. He, "Cm-nas: Cross-modality neural architecture search for visible-infrared person re-identification," in *ICCV*, 2021. 6, 7
- [52] D. Nguyen, H. Hong, K. Kim, and K. Park, "Person recognition system based on a combination of body images from visible light and thermal cameras," *Sensors*, 2017. 5, 6
- [53] Y.-B. Zhao, J.-W. Lin, Q. Xuan, and X. Xi, "Hpiln: a feature learning framework for cross-modality person re-identification," *IET Image Processing*, 2019. 6, 7
- [54] Y. Ling, Z. Luo, Y. Lin, and S. Li, "A multi-constraint similarity learning with adaptive weighting for visible-thermal person re-identification," in *IJCAI*, 2021. 7
- [55] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *JMLR*, 2008. 8, 9
- [56] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *ECCV*, 2016. 8, 9
- [57] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017. 8, 9