Heterogeneous Test-Time Training for Multi-Modal Person Re-identification

Zi Wang[1], Huaibo Huang[2], Aihua Zheng[3]*, Ran He[2]

[1]School of Computer Science and Technology, Anhui University, Hefei, China
[2]MAIS & CRIPAC, CASIA, Beijing, China
[3]Information Materials and Intelligent Sensing Laboratory of Anhui Province, Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, School of Artificial Intelligence, Anhui University, Hefei, China
ziwang1121@foxmail.com, huaibo.huang@cripac.ia.ac.cn, ahzheng214@foxmail.com, rhe@nlpr.ia.ac.cn

Abstract

Multi-modal person re-identification (ReID) seeks to mitigate challenging lighting conditions by incorporating diverse modalities. Most existing multi-modal ReID methods concentrate on leveraging complementary multi-modal information via fusion or interaction. However, the relationships among heterogeneous modalities and the domain traits of unlabeled test data are rarely explored. In this paper, we propose a **H**eterogeneous **T**est-time **T**raining (**HTT**) framework for multi-modal person ReID. We first propose a Cross-identity Inter-modal Margin (CIM) loss to amplify the differentiation among distinct identity samples. Moreover, we design a Multi-modal Test-time Training (MTT) strategy to enhance the generalization of the model by leveraging the relationships in the heterogeneous modalities and the information existing in the test data. Specifically, in the training stage, we utilize the CIM loss to further enlarge the distance between anchor and negative by forcing the inter-modal distance to maintain the margin, resulting in an enhancement of the discriminative capacity of the ultimate descriptor. Subsequently, since the test data contains characteristics of the target domain, we adapt the MTT strategy to optimize the network before the inference by using self-supervised tasks designed based on relationships among modalities. Experimental results on benchmark multi-modal ReID datasets RGBNT201, Market1501-MM, RGBN300, and RGBNT100 validate the effectiveness of the proposed method. The codes can be found at https://github.com/ziwang1121/HTT.

Introduction

The task of multi-modal person re-identification (ReID) has attracted increasing attention as a result of the advancement of machine learning technology and the concentration of social security issues. Different from the primary objectives of single-modal ReID methods and cross-modal ReID methods, multi-modal ReID aims to alleviate the issue of insufficient information in single-modal ReID and spectral disparity in cross-modal ReID. Zheng et al. (Zheng et al. 2021) propose the multi-modal person ReID, introducing complementary information among various modalities. The benchmark multi-modal dataset is comprised of data from visible (RGB), near-infrared (NI), and thermal infrared (TI).

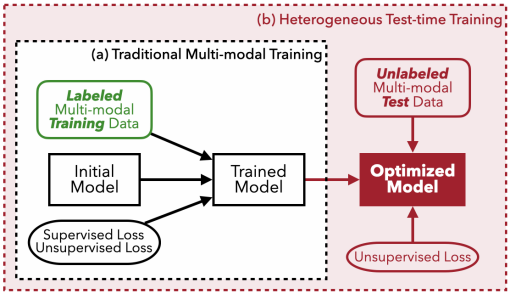

Figure 1: (a) Traditional multi-modal training exclusively utilizes labeled training data. (b) The proposed heterogeneous test-time training additionally leverages unlabeled test data for optimization.

The majority of multi-modal ReID methods (Zheng et al. 2021; Li et al. 2020; Wang et al. 2022d; Guo et al. 2022) place more emphasis on the fusion or interaction of modalities. The current multi-modal ReID methods can be categorized into traditional multi-modal training, using only labeled training data, and trained under the constraints of supervised and unsupervised loss, as shown in Fig. 1 (a).

However, there are two common but rarely explored clues in the multi-modal person ReID. **(1) The relationships between heterogeneous modalities.** The multi-modal ReID methods employ various losses to supervise network learning. Cross-entropy (CE) loss can force the network to distinguish different identities, as shown in Fig. 2 (a). Fig. 2 (b) illustrates that after calculating the distance between the anchor and the positive/negative, triplet loss (Schroff, Kalenichenko, and Philbin 2015) ensures the difference between them is larger than the predetermined margin. As shown in Fig. 2 (c), the multi-modal margin (3M) loss (Wang et al. 2022d) is designed to increase the distance among the intra-identity modalities. Nevertheless, the distinction between the final person descriptors is predominantly governed by inter-modal disparities. The discriminative potency of the final descriptor can be enhanced by constraining inter-modal relationships, which are pivotal yet disregarded by these methodologies. **(2) The unlabeled multi-modal test data.** Samples in the training and test sets possess distinct identities and might originate from diverse

*Corresponding Author.

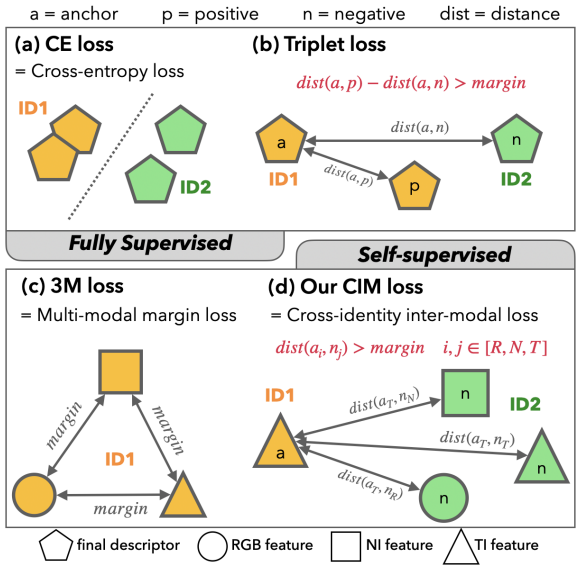

Figure 2: Comparison of diverse losses. CE loss separates features from two identities. Triplet loss controls the distance of features. 3M loss constrains the distance among modalities in each sample. CIM loss constrains the distance between inter-modal features from different identities.

and intricate environments. As a result, the classifier is unavailable, and the style shifts are unpredictable in the test phase. However, the unlabeled multi-modal test data comprises abundant domain-specific information and available modal relationships. Most multi-modal ReID methods neglect to exploit the relationships within unlabeled multi-modal test data and design appropriate self-supervised tasks for test-time training.

To enhance model generalization through inter-modal relationship mining and test data utilization, we propose the Heterogeneous Test-time Training (HTT) framework for robust multi-modal person ReID. First, we design the cross-identity inter-modal margin (CIM) loss to enhance distinctiveness among samples. In particular, we calculate the inter-modal distances between the anchor and the negative, then restrict the distances to surpass a specified margin value, as shown in Fig. 2 (d). The distinction across heterogeneous modal features can be enlarged by employing our CIM loss, which will ultimately increase the discriminative ability of person descriptors. Additionally, we propose the multi-modal test-time training (MTT) strategy, which capitalizes on unlabeled test data before inference. During the test-time training stage, two self-supervised tasks that rely on multi-modal data are constructed and used to optimize the trained model. On the one hand, we employ the 3M loss (Wang et al. 2022d) to quantify the intra-sample discrepancy. On the other hand, we utilize the proposed CIM loss to impose constraints on the inter-modal distance among distinct identities. Our model can be trained on the source data as conventional ReID. Moreover, the pre-trained model will be further optimized exclusively through self-supervised tasks tailored for unlabeled test data, as shown in Fig. 1 (b). We provide sufficient experimental results and in-depth analyses to show the advantages of the proposed method. Here are our main contributions:

- We propose a novel method for the multi-modal person ReID task, termed heterogeneous test-time training (HTT), to improve performance on unseen test data by utilizing the relationship between heterogeneous modalities and fine-tuning the network before inference.
- We introduce the cross-identity inter-modal margin (CIM) loss to further enhance the discriminant of the final descriptor by measuring the inter-modal distance between the anchor and the negative samples and constraining the distance to be larger than the preset margin.
- We design the multi-modal test-time training (MTT) strategy to enhance the generalization of the model on unseen test data that contains domain characteristics and modal relationships through two self-supervised tasks.
- We employ extensive ablation studies and experimental comparisons against state-of-the-art approaches on the four standard benchmark multi-modal ReID datasets to demonstrate the effectiveness of our method.

Related Work

Multi-modal Re-identification

In order to alleviate the problem that single-modal visible light data cannot provide useful information at night(Zhu et al. 2021; Rao et al. 2021; Wang et al. 2022a; Li et al. 2021; Li, Wu, and Zheng 2021; Zheng et al. 2015; Zhou et al. 2023; Zhang et al. 2023; Li et al. 2022; Lei et al. 2008), and eliminate the huge domain heterogeneity between cross-modal data (Zhang et al. 2022; Tian et al. 2021; Wu et al. 2021; Chen et al. 2021; Wu et al. 2017; Nguyen et al. 2017; Huang et al. 2022; Farooq et al. 2022; Kim et al. 2023; Feng, Wu, and Zheng 2023; Zhang and Wang 2023; Zheng et al. 2023; He et al. 2015), Zheng et al. (Zheng et al. 2021) construct the first multi-modal person ReID dataset, RGBNT201. And Zheng et al. (Li et al. 2020) propose the benchmark multi-modal vehicle ReID datasets, RGBN300 and RGBNT100. These multi-modal datasets proposed for the ReID task all contain complementary data from multiple modalities (visible, near-infrared, and thermal infrared). These complementary multi-modal images provide new solutions to the traditional ReID challenges but also bring additional issues existing in heterogeneous data. To merge the useful information from multi-modal data, many novel approaches have emerged. PFNet (Zheng et al. 2021) proposes to extract the modal features by the multi-branch network without sharing parameters, then divide the global features into several parts, and finally fuse the features at both the global and local levels. IEEE (Wang et al. 2022d) proposes to learn more specific information for each modality by introducing cross-modal interaction and multi-modal margin loss and leveraging local details by designing the relation-based enhancement module. HAMNet (Li et al. 2020) provides a powerful baseline framework for multi-modal vehicle ReID by automatically fusing spectrum-specific features in the network and introducing heterogeneity-collaboration

loss. GAFNet (Guo et al. 2022) designs the input-level generated transitional modality model to involve different data distributions and introduces a feature-level attentive module to fuse different modalities. All the above-mentioned methods consider the fusion or interaction of information. However, the deeper modal relationships in training data and the domain characteristics in test data are rarely explored.

Test-time Training Strategy

To improve the generalization of the model trained only in the source domain on the target set with large domain shifts, Sun et al. (Sun et al. 2020) first propose test-time training for enhancing the generalization of the trained modal. In (Sun et al. 2020), the model is optimized by both the self-supervised loss and the main task loss during training, but only the self-supervised loss is employed to fine-tune the network in test-time training. This general approach proposed in (Sun et al. 2020) is easy to combine with other tasks and effective for them. Many researchers introduce the test-time training to the downstream tasks to improve the generalization of the model on the out-of-distribution test data (Han et al. 2022; Wang et al. 2020; Shin et al. 2022; Liu et al. 2022; Gandelsman et al. 2022). Shin et al. (Shin et al. 2022) propose two complementary modules, intra-modal pseudo-label generation, and inter-modal pseudo-label refinement, to take full advantage of self-supervising signals provided by multi-modality. Han et al. (Han et al. 2022) propose a test-time training ReID framework to update BN parameters adaptively by two designed self-supervised tasks. However, the self-supervised modules in (Shin et al. 2022) are specially designed for point clouds and RGB data. The self-supervised tasks proposed in (Han et al. 2022) do not consider the advantages existing in multi-modal data and introduce additional networks and modules to assist prediction. Therefore, these test-time training strategies cannot be directly applied to multi-modal ReID tasks.

Method

In this section, we will introduce our proposed method in detail. We start by describing the baseline framework. Then we illustrate more details about the proposed Heterogeneous Test-time Training (HTT), including two key components: (1) Cross-identity Inter-modal Margin (CIM) Loss, which constrains the modal feature distances among different samples during training. (2) Multi-modal Test-time Training (MTT) strategy, which further fine-tunes the model by employing multi-modal margin loss and designed CIM loss.

The Baseline Framework

In the training phase of the multi-modal person ReID task, each sample sent to the backbone is a triplet composed of three aligned images from visible (R), near-infrared (N), and thermal infrared (T). The $input$ can be denoted as:

$$input = [I_R^{sou}, I_N^{sou}, I_T^{sou}], \tag{1}$$

where $[I_R^{sou}, I_N^{sou}, I_T^{sou}]$ represents the image triplet in the source domain for training, and images in the triplet have the same size with $[256, 128, 3]$ in [height, width, channel]. Due to the excellent performance of single-modal ReID methods based on vision transformers (Wang et al. 2022b,c,e; He et al. 2021), we choose the basic vision transformer (Dosovitskiy et al. 2020) as the feature extractor in our method. The corresponding three modal features of the source are obtained after the ViT-based backbone. The process of feature extraction can be formulated as follows:

$$[f_R^{sou}, f_N^{sou}, f_T^{sou}] = \phi\left([I_R^{sou}, I_N^{sou}, I_T^{sou}]\right), \tag{2}$$

where ϕ denotes the ViT-based backbone, and $[f_R^{sou}, f_N^{sou}, f_T^{sou}]$ denotes the modal features of the corresponding modality and is used for multi-modal margin loss (Wang et al. 2022d) and proposed CIM loss. The dimension of each modal feature is 768-dim.

Three modal features are combined to form the global feature by concatenating. Then the global feature passes through the normalization layer to obtain the normalized feature. Finally, we employ the classifier layer to predict the identity of each sample:

$$\begin{aligned} f_g^{sou} &= C([f_R^{sou}, f_N^{sou}, f_T^{sou}], dim = 1), \\ f_{BN}^{sou} &= BN(f_g^{sou}), \end{aligned} \tag{3}$$

$$p = cls(f_{BN}^{sou}), \tag{4}$$

where $C(\cdot)$ denotes the concatenation operation employed on the channel-wise of modal features. f_g^{sou} represents the global feature, which is used for computing triplet loss. $BN(\cdot)$ denotes the batch normalization layer. cls denotes the classifier layer. p represents the identity prediction result used for cross-entropy loss. f_{BN}^{sou} denotes the normalized feature, which is the final descriptor of the sample. Compared with the training process, the process of the test phase lacks the step of identity prediction, and the f_{BN}^{tar} obtained from $[I_R^{tar}, I_N^{tar}, I_T^{tar}]$ in the target domain (test set) will be used for the final evaluation.

Cross-identity Inter-modal Margin Loss

To further enlarge the distinction between the anchor and the negative samples, we design the cross-identity inter-modal margin (CIM) loss for multi-modal person ReID. It is worth noting that the CIM loss is essentially self-supervised and can be utilized in the training and testing stages. For each iteration of the training phase, the data loader selects several samples for the batch. The specially designed data loader can ensure that each sample in the current batch has negative samples from different identities. We first extract the modal features from both the anchor and the negative, as follows:

$$\begin{aligned} [f_R^a, f_N^a, f_T^a] &= \phi(input^a), \\ [f_R^n, f_N^n, f_T^n] &= \phi(input^n), \end{aligned} \tag{5}$$

where a and n indicate the feature comes from the anchor and the negative samples, respectively. After feature extraction, we can compute all the distances between cross-identity modalities. Note that to simplify the notation, all symbols omit the superscript ^{sou} indicating the source domain. **For each anchor/negative pair**, we choose the inter-modal distance D_{im} to represent the discrepancy between

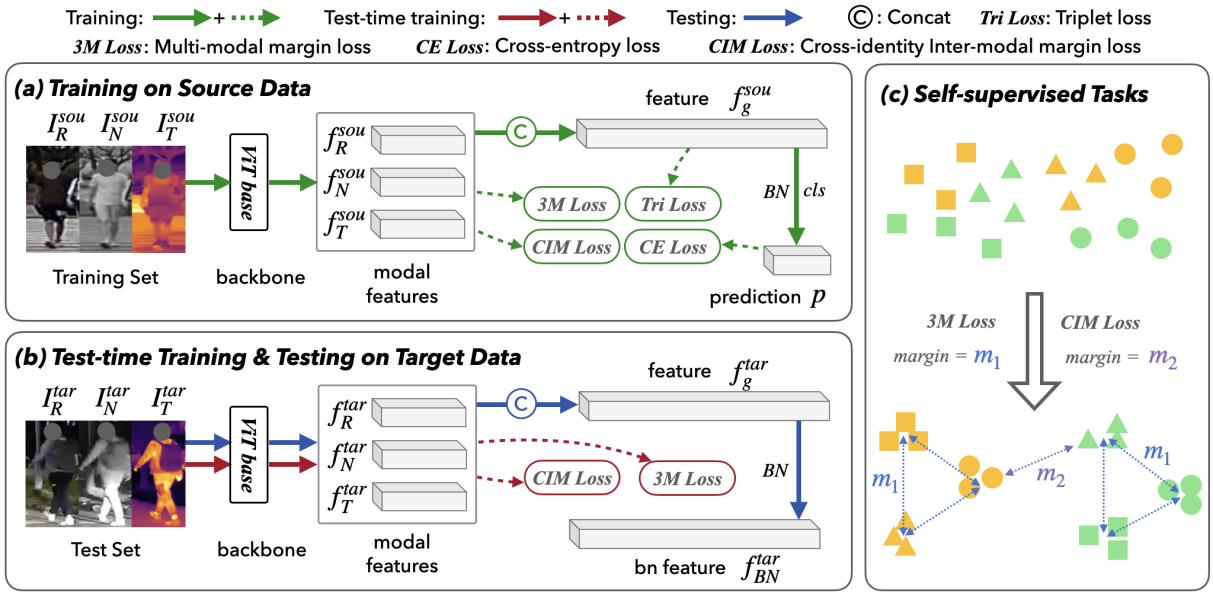

Figure 3: Overview of the proposed Heterogeneous Test-time Training (HTT) framework. (a) To constrain the learning of the model during training, we employ the combination of two fully supervised losses (CE loss and Tri loss) and two self-supervised losses (3M loss and CIM loss). (b) Only the self-supervised losses are employed to update the model during test-time training. After fine-tuning, the normalized feature f_{BN}^{tar} will be used for testing. (c) The 3M loss increases the distance among the intra-identity modalities. The CIM loss further enlarges the distinction of modal features belonging to different identities.

them.

$$D_{im}(f^a, f^n) = max(|margin_{_{CIM}} - dist_m(f_{z_1}^a, f_{z_2}^n)|), \quad z_1, z_2 \in (R,N,T), \tag{6}$$

where z_1 and z_2 denote the modality of the feature. $margin_{_{CIM}}$ denotes the predetermined margin between cross-identity modalities. $dist_m$ denotes the Manhattan distance, which can be computed as follows:

$$dist_m(f_1, f_2) = \sum_{i=1}^{K} |f_1^i - f_2^i|, \tag{7}$$

where K denotes the dimension of the feature f. **For each batch**, we choose the average of all inter-sample distances as the value of CIM loss.

$$\ell_{CIM} = \frac{1}{B} \sum_{i=1}^{B} D_{im}(a_i, n_i), \tag{8}$$

where B denotes the batch size of each iteration. Based on these two criteria, in each iteration, we use the distance of the two modal features from different identity samples to optimize the network. The proposed CIM loss compels the network to prioritize inter-modal disparities among samples from distinct identities, thereby enhancing the diversity of ultimate features.

Multi-modal Test-time Training Strategy

To improve the generalization capabilities of the trained model, we introduce the multi-modal test-time training (MTT) strategy, which leverages unlabeled data from the target domain. MTT follows the setting of basic test-time training (Sun et al. 2020), only the labeled source data is used for training, and the unlabeled target data is used for fine-tuning the trained model. During the test-time training phase, we utilize two self-supervised loss functions: the multi-modal margin (3M) loss (Wang et al. 2022d) and the cross-identity inter-modal margin (CIM) loss. The comprehensive procedure of MTT is depicted in Algorithm. 1.

Initially, we build a subset D' consisting of randomly selected samples from the unlabeled test sets D for MTT. In each batch, we first calculate the 3M loss for each sample using Eq.11, as the constraint is exclusively applied within individual sample triplets. Subsequently, we compute the CIM loss between distinct samples. Nevertheless, during the testing phase, the genuine labels of the inputs remain unavailable, rendering it impossible to ascertain the consistency of the two sample identities. To address this issue and successfully utilize CIM loss, we employ a straightforward yet logical approach: employing the trained classifier to assign pseudo-labels to the sampled data. If the pseudo-labels of samples exhibit inconsistency, we posit that they pertain to positive and negative sample pairs and calculate the CIM loss according to Eq.8. Otherwise, the CIM loss is designated as zero. The ultimate loss $\mathcal{L}_{TTT}$ during the test-time training phase is the aggregate of the 3M loss and the CIM loss. The pre-trained model will undergo additional optimization through the application of $\mathcal{L}_{TTT}$. Importantly, the batch size should be no less than 2 to enable the utilization of CIM loss. This necessity arises from the fact that the CIM loss requires the computation of inter-modal distances be-

Algorithm 1: Multi-modal Test-time Training Strategy

Require: Parameter θ of the trained model, batch size B, learning rate λ, test dataset D

while not done **do**
  Select the subset $D' \subseteq D$ for test-time training
  Randomly select B samples from D'
  for $i = 1$ **to** B **do**
    Compute the 3M loss ℓ_{3M} via Eq. (11)
    for $j = i$ **to** B **do**
      Predict the pseudo labels p_i and p_j via Eq. (4)
      if $p_i \neq p_j$ **then**
        Compute the CIM loss ℓ_{CIM} via Eq. (8)
      else
        $\ell_{CIM} = 0$
      end if
    end for
    Evaluate the loss $\mathcal{L}_{TTT}$ via Eq. (13)
  end for
  Optimize the model parameter θ with SGD:
  $\theta \leftarrow \theta - \lambda \cdot \nabla_\theta \frac{1}{B} \sum_{i=1}^{B} \mathcal{L}_{TTT}$
end while

Return: Updated the whole model

tween the anchor and negative samples.

Loss Functions

There are four varieties of losses used in the proposed method, including cross-entropy (CE) loss, triplet loss, multi-modal margin (3M) loss, and the proposed cross-identity inter-modal margin (CIM) loss. As shown in Fig. 3, in the training phase, we calculate the 3M loss and CIM loss on the modal features extracted by the ViT-based backbone, compute the Tri loss on the global feature concatenated by modal features, and employ the CE loss on the prediction result output by the classifier layer. During the test-time training, only the 3M loss and the CIM loss are used for self-supervised learning. The widely used CE loss measure, ℓ_{CE}, can assist the network in distinguishing samples that belong to different identities:

$$\ell_{CE} = -\frac{1}{B} \sum_{i=1}^{B} \sum_{j=1}^{M} y_{ij} \log\left(p_{ij}\right), \tag{9}$$

where B denotes the number of samples in each batch, y_{ij} denotes the ground truth of each sample, M denotes the total identity of the person in the training set, and p_{ij} represents the probability that sample i belongs to identity j. The triplet loss is always employed to ensure the distance between the anchor and the negative is larger than the predetermined margin. ℓ_{Tri} can be computed as follows:

$$\ell_{Tri} = \sum_{i=1}^{P} \sum_{a=1}^{K} [m + \overbrace{\max_{p=1,\ldots,K} Dis(f_a^i, f_p^i)}^{\text{hardest positive}} - \overbrace{\min_{n=1,\ldots,K} Dis(f_a^i, f_n^i)}^{\text{hardest negative}}], \tag{10}$$

where f_a is the anchor feature, f_p is the positive feature with the same identity as f_a, and f_n is the negative feature with a different identity in the batch. The multi-modal margin loss is employed to ensure the intra-sample inter-modal distance maintains the predetermined margin. The multi-modal margin loss ℓ_{3M} for each sample can be formulated as follows:

$$\begin{aligned} \ell_{3M} &= max(m_{_{3M}} - \|f_i - f_j\|_2^2) \\ & i, j \in [R, N, T], \quad i \neq j \end{aligned} \tag{11}$$

where $m_{_{3M}}$ indicates the predetermined distance among intra-sample modalities. Moreover, we propose the CIM loss to expand the distinction between the anchor and the negative samples.

The final losses $\mathcal{L}_{train}$ and $\mathcal{L}_{TTT}$ used for training and test-time training can be formulated as follows:

$$\mathcal{L}_{train} = \ell_{CE} + \ell_{Tri} + \alpha_1 * \ell_{3M} + \beta_1 * \ell_{CIM}, \tag{12}$$

$$\mathcal{L}_{TTT} = \alpha_2 * \ell_{3M} + \beta_2 * \ell_{CIM}, \tag{13}$$

where α_1 and β_1 denote the balancing hyperparameters for ℓ_{3M} and ℓ_{CIM} in training. α_2 and β_2 are hyperparameters to balance the self-supervised losses.

Experiments

Datasets and Evaluation Protocols

We first introduce multi-modal person ReID datasets RGBNT201 (Zheng et al. 2021) and Market1501-MM (Wang et al. 2022d), two multi-modal vehicle ReID datasets RGBNT100 and RGBN300 built by (Li et al. 2020). Then we illustrate the evaluation protocols used in our test phase. **Datasets**. (1) RGBNT201, the first multi-modal person ReID dataset, which contains 4787 image triplets of 201 identities. Each image triplet consists of three aligned images: visible, near-infrared, and thermal infrared. (2) Market1501-MM, all near-infrared and thermal-infrared images are generated from visible images by pre-trained cycle-GAN(Zhu et al. 2017). (3) RGBN300, the dual-modal ReID dataset, which contains 50125 image pairs of 300 vehicle identities captured by visible and near-infrared cameras. (4) RGBNT100 is extended based on RGBN300. Additional captured 17250 thermal images and corresponding visible and near-infrared image pairs constitute the dataset with 17250 image triples.

Evaluation Protocols. Following conventions in the ReID community, we employ the mean average precision (mAP) and cumulative matching characteristic curve (CMC) to evaluate the performance of the proposed method and other methods on standard datasets. According to the Euclidean distance, CMC scores reflect the retrieval precision, and the rank-n indicates the first n samples with the same identity from different cameras that are closest to the query.

Implementation Details

The implementation platform of our method is Pytorch (Paszke et al. 2019) with one RTX 3090Ti GPU. We use the basic vision transformer (Dosovitskiy et al. 2020) with stride=16 as the backbone for feature extraction. All the images are resized to 256×128. All training images are

Methods		*RGBNT201*				*Market1501-MM*			
		mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10
Single-modal	MLFN (CVPR2018)	24.7	23.7	38.5	49.5	42.7	68.1	87.1	92.0
	HACNN (CVPR2018)	19.3	14.7	25.5	32.8	42.9	69.1	86.6	92.2
	OSNet (ICCV2019)	22.1	22.9	37.2	45.9	39.7	69.3	86.7	91.3
Multi-modal	HAMNet (AAAI2020)	27.7	26.3	41.5	51.7	60.0	82.8	92.5	95.0
	PFNet (AAAI2021)	38.5	38.9	52.0	58.4	60.9	83.6	92.8	95.5
	IEEE (AAAI2022)	46.4	47.1	58.5	64.2	64.3	**83.9**	93.0	95.7
	HTT w/o MTT	69.0	70.0	80.5	85.6	65.9	80.3	95.4	97.6
	HTT (Ours)	**71.1**	**73.4**	**83.1**	**87.3**	**67.2**	81.5	**95.8**	**97.8**

Table 1: Results of our method on RGBNT201 and Market1501-MM compared with state-of-the-art methods (in %).

augmented before being sent to the backbone, $i.e.$, random erasure, random flipping, and padding following (He et al. 2021). The Stochastic Gradient Descent (SGD) (Bottou 2012) with a weight decay of 0.0001 is used in our experiment to fine-tune the model for both training and test-time training. **Training phase**. The learning rate in the training phase is set to 0.008. The maximum epoch is 80. The batch size is set to 32, consisting of 32 image triplets from four different identities. The dimension of each modal feature is 768-dim, while the global feature and the normalized feature have the dimension of 2304-dim. The weights of (α_1, β_1) are (0.5, 0.5). **Test time training phase**. The learning rate for test-time training is 0.001. And we fine-tune the network once by only using self-supervised losses ℓ_{3M} and ℓ_{CIM}. The batch size is set to 16, consisting of randomly selected image triplets from unlabeled test data. The balancing hyperparameters (α_2, β_2) are (1, 1) for RGBNT201 dataset.

Comparison with State-of-the-art Methods

We first evaluate the effectiveness of the proposed method on the standard multi-modal person ReID datasets. Additionally, we also conduct experiments on multi-modal vehicle ReID datasets, which also demonstrate the superiority and adaptability of our method.

Experiments on person ReID datasets. (Zheng et al. 2021). We compare the proposed method with the state-of-the-art single-modal and multi-modal person ReID methods on RGBNT201 and Market1501-MM, including MLFN (Chang, Hospedales, and Xiang 2018), HACNN (Li, Zhu, and Gong 2018), OSNet (Zhou et al. 2019), HAMNet (Li et al. 2020), PFNet (Zheng et al. 2021) and IEEE (Wang et al. 2022d). As shown in Table 1, when testing on the RGBNT201 dataset, our method HTT without multi-modal test-time training (MTT) achieves 69.0% mAP and 70.0% on Rank-1 accuracy, all the results outperform the state-of-the-art methods. Due to the powerful feature extraction capability of the ViT-based backbone and the strong constraints of designing CIM loss, our method outperforms other methods by 22.6% and 22.9% in mAP and Rank-1 at least. Moreover, we further evaluate the performance of employing the MTT strategy. As shown in the last line, the full HTT achieves the best results on all evaluation protocols with 71.1% mAP and 73.4% on Rank-1 accuracy. Specifically, the mAP and Rank-1 have increased by 2.1% and 3.4%. These improvements in the indicators prove that employing MTT on unlabeled test data is effective for multi-modal ReID. Meanwhile, on the Market1501 dataset, our HTT achieves the highest mAP 67.2%, which is 2.9% higher than the second place. At the same time, our Rank-5 and Rank-10 are also the highest among all methods, which are 95.8% and 97.8% respectively.

Methods	*RGBN300*		*RGBNT100*	
	mAP	R-1	mAP	R-1
PCB (ECCV2018)	57.7	82.0	57.2	83.5
MGN (ACM MM2018)	60.5	83.7	58.1	83.1
ABD (ICCV2019)	58.9	83.1	60.4	85.1
HAMNet (AAAI2020)	61.9	84.0	64.1	84.7
GAFNet (ICSP2022)	72.7	**91.9**	74.4	**93.4**
HTT (Ours)	**77.1**	90.8	**75.7**	92.6

Table 2: Experimental results of our method on multi-modal vehicle ReID datasets RGBN300 and RGBNT100 compared with state-of-the-art methods (in %).

Experiments on vehicle ReID datasets. We evaluate our method on RGBN300 and RGBNT100, compared with the state-of-the-art single-modal and multi-modal vehicle ReID methods, including PCB (Sun et al. 2018), MGN (Wang et al. 2018), ABD (Chen et al. 2019), HAMNet (Li et al. 2020), GAFNet (Guo et al. 2022). As shown in Table 2, when testing on the RGBN300, the full HTT achieves the best results with 77.1% mAP and the second-best results with 90.8% on Rank-1 accuracy. Our mAP results have exceeded GAFNet by 4.4% and are only 1.1% lower in Rank-1. When testing on RGBNT100, we achieve the best results with 75.7% mAP and the second-best results with 92.6% on Rank-1 accuracy. It is worth noting that the generator used in GAFNet relies on training data. On the contrary, our method can achieve results close to or surpassing SOTA without much adjustment, which verifies adaptability.

Ablation Study

We further conduct ablation studies on the RGBNT201 dataset to analyze the effectiveness of two key components

	B	CIM	MTT	mAP	R-1	R-5	R-10
(1)	✓	-	-	67.5	69.4	81.2	85.0
(2)	✓	✓	-	69.0	70.3	81.5	85.6
(3)	✓	✓	✓	**71.1**	**73.4**	**83.1**	**87.3**

Table 3: Ablation study of proposed cross-identity inter-modal margin (CIM) loss and multi-modal test-time training (MMT) strategy on the baseline (B) on RGBNT201 (in %).

in our HTT: cross-identity inter-modal margin (CIM) loss and multi-modal test-time training strategy (MTT).

The experimental results of the ablation study are shown in Table 3. We first apply the proposed CIM loss to the baseline for comparison, as shown in lines (1) and (2). The model will be more discriminative for different identities because our CIM loss can further widen the gap between various samples. As a result, after using CIM, mAP and Rank-1 are increased by 1.5% and 0.9%, and Rank-5 and Rank-10 are also increased by 0.3% and 0.6%, respectively. From the comparison of the results in lines (1) and (3), the highest results can be obtained by using both the proposed CIM and MTT, and the mAP and Rank-1 values are improved by 3.6% and 4.0%, respectively. The above results verify the effectiveness of our proposed CIM loss in the training and test-time training phases, as well as the effectiveness of the MMT strategy under self-supervised loss constraints. With the help of CIM and MTT, our model can achieve the best experimental results on multi-modal test data with unknown domain shifts.

Discussion on CIM Loss

Our proposed CIM loss is an unsupervised loss, independent of network structure or data annotations, rendering it applicable to diverse ReID methods. We selected representative methods from single-modal and multi-modal approaches, OSNet (Zhou et al. 2019) and IEEE(Wang et al. 2022d), and incorporated the proposed CIM loss into their training process. As shown in Table 4, when combined with OSNet, the mAP and Rank-1 accuracy increase by 1.4% and 0.2% respectively. Furthermore, when integrated with IEEE, the method specifically designed for multi-modal ReID, the mAP and Rank-1 improvements reach as high as 2.8% and 1.4% respectively. The outcomes resulting from the amalgamation of CIM loss with these two methods demonstrate its remarkable transferability.

Methods	mAP	R-1	R-5	R-10
OSNet* (ICCV2019)	23.4	23.7	39.2	48.7
+ CIM loss	**24.8**	**23.9**	**44.3**	**55.5**
IEEE* (AAAI2022)	46.5	47.8	59.2	64.8
+ CIM loss	**49.3**	**49.2**	**62.4**	**69.4**

Table 4: The results of combining CIM loss with different methods on RGBNT201 (in %). * indicates the results are reproduced by us.

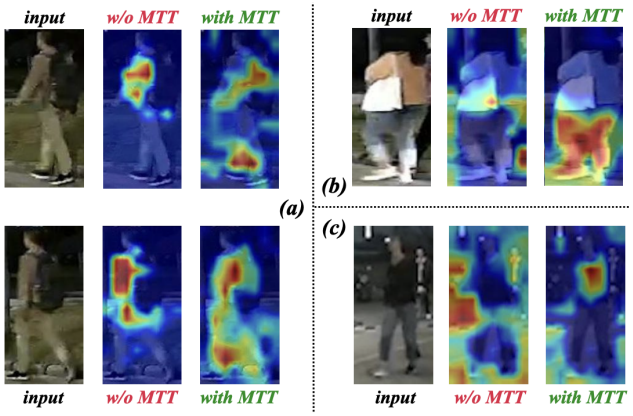

Figure 4: Visualization results without and with multi-modal test-time training on unseen test data, drawn by Gradient-weighted Class Activation Mapping (Grad-CAM).

Discussion on MTT Strategy

To demonstrate the significance of the MTT strategy on unseen data in a more intuitive manner, we further provide visualization results by using Gradient-weighted Class Activation Mapping (Grad-CAM) (Selvaraju et al. 2017).

Illustrated in Fig. 4 (a), upon employing MTT, the model exhibits an extended capability to concentrate on a wider range of bodily regions, thereby enhancing the overall information density of features. For some samples, the model devoid of MTT fails to allocate attention toward the individual's region or, in some cases, erroneously fixates on the background. After the implementation of MTT, the model discerns salient regions, underscoring the capability of MTT to enhance the model's generalization and adaptability to uncharted test samples, as depicted in Fig. 4 (b) and (c).

Conclusion

This paper presents an extensive study of the test-time training strategy for multi-modal person ReID, with the following contributions: (1) We propose a heterogeneous test-time training **(HTT)** framework to enhance the generalization of the trained model on unseen test data by fine-tuning before inference. (2) We design a self-supervised loss, cross-identity inter-modal margin **(CIM)** loss, to enhance the discriminant of the descriptor by constraining the inter-modal distance between the anchor and the negative. (3) We combine the CIM loss with other self-supervised losses for multi-modal test-time training **(MTT)** to adapt the model to the unlabeled test data. Extensive experimental results demonstrate the effectiveness of the proposed method on several multi-modal ReID datasets. Our approach attains state-of-the-art performance while necessitating merely uncomplicated fine-tuning employing unlabeled test data. In our future work, we will persist in exploring methods to ameliorate the quality of sampled data and dynamically update the network architecture.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (Grants 62372003), the Natural Science Foundation of Anhui Province (Grants 2308085Y40), the University Synergy Innovation Program of Anhui Province (Grant GXXT-2022-036), the National Natural Science Foundation of China (Grant No. 62006228), the Youth Innovation Promotion Association CAS (Grant No. 2022132), and Beijing Nova Program (20230484276).

References

Bottou, L. 2012. Stochastic gradient descent tricks. In *Neural networks: Tricks of the trade*, 421–436. Springer.

Chang, X.; Hospedales, T. M.; and Xiang, T. 2018. Multi-level factorisation net for person re-identification. In *CVPR*, 2109–2118.

Chen, T.; Ding, S.; Xie, J.; Yuan, Y.; Chen, W.; Yang, Y.; Ren, Z.; and Wang, Z. 2019. Abd-net: Attentive but diverse person re-identification. In *ICCV*, 8351–8361.

Chen, Y.; Wan, L.; Li, Z.; Jing, Q.; and Sun, Z. 2021. Neural feature search for rgb-infrared person re-identification. In *CVPR*, 587–597.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Farooq, A.; Awais, M.; Kittler, J.; and Khalid, S. S. 2022. AXM-Net: Implicit cross-modal feature alignment for person re-identification. In *AAAI*, volume 36, 4477–4485.

Feng, J.; Wu, A.; and Zheng, W.-S. 2023. Shape-Erased Feature Learning for Visible-Infrared Person Re-Identification. In *CVPR*, 22752–22761.

Gandelsman, Y.; Sun, Y.; Chen, X.; and Efros, A. A. 2022. Test-time training with masked autoencoders. In *NeurIPS*, volume 35, 29374–29385.

Guo, J.; Zhang, X.; Liu, Z.; and Wang, Y. 2022. Generative and Attentive Fusion for Multi-spectral Vehicle Re-Identification. In *ICSP*, 1565–1572. IEEE.

Han, K.; Si, C.; Huang, Y.; Wang, L.; and Tan, T. 2022. Generalizable person re-identification via self-supervised batch norm test-time adaption. In *AAAI*, volume 36, 817–825.

He, R.; Zhang, M.; Wang, L.; Ji, Y.; and Yin, Q. 2015. Cross-modal subspace learning via pairwise constraints. *TIP*, 24: 5543–5556.

He, S.; Luo, H.; Wang, P.; Wang, F.; Li, H.; and Jiang, W. 2021. Transreid: Transformer-based object re-identification. In *ICCV*, 15013–15022.

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*, volume 36, 1034–1042.

Kim, M.; Kim, S.; Park, J.; Park, S.; and Sohn, K. 2023. PartMix: Regularization Strategy to Learn Part Discovery for Visible-Infrared Person Re-identification. In *CVPR*, 18621–18632.

Lei, Z.; Liao, S.; He, R.; Pietikainen, M.; and Li, S. Z. 2008. Gabor volume based local binary pattern for face representation and recognition. In *FG*, 1–6.

Li, H.; Li, C.; Zhu, X.; Zheng, A.; and Luo, B. 2020. Multi-spectral vehicle re-identification: A challenge. In *AAAI*, volume 34, 11345–11353.

Li, H.; Wu, G.; and Zheng, W.-S. 2021. Combined depth space based architecture search for person re-identification. In *CVPR*, 6729–6738.

Li, H.; Ye, M.; Wang, C.; and Du, B. 2022. Pyramidal transformer with conv-patchify for person re-identification. In *ACM MM*, 7317–7326.

Li, W.; Zhu, X.; and Gong, S. 2018. Harmonious attention network for person re-identification. In *CVPR*, 2285–2294.

Li, Y.; He, J.; Zhang, T.; Liu, X.; Zhang, Y.; and Wu, F. 2021. Diverse part discovery: Occluded person re-identification with part-aware transformer. In *CVPR*, 2898–2907.

Liu, Q.; Chen, C.; Dou, Q.; and Heng, P.-A. 2022. Single-domain generalization in medical image segmentation via test-time adaptation from shape dictionary. In *AAAI*, volume 36, 1756–1764.

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

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 32.

Rao, Y.; Chen, G.; Lu, J.; and Zhou, J. 2021. Counterfactual attention learning for fine-grained visual categorization and re-identification. In *ICCV*, 1025–1034.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 815–823.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 618–626.

Shin, I.; Tsai, Y.-H.; Zhuang, B.; Schulter, S.; Liu, B.; Garg, S.; Kweon, I. S.; and Yoon, K.-J. 2022. MM-TTA: Multi-Modal Test-Time Adaptation for 3D Semantic Segmentation. In *CVPR*, 16928–16937.

Sun, Y.; Wang, X.; Liu, Z.; Miller, J.; Efros, A.; and Hardt, M. 2020. Test-time training with self-supervision for generalization under distribution shifts. In *ICML*, 9229–9248.

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*, 480–496.

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*, 1522–1531.

Wang, D.; Shelhamer, E.; Liu, S.; Olshausen, B.; and Darrell, T. 2020. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*.

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*, 274–282.

Wang, H.; Shen, J.; Liu, Y.; Gao, Y.; and Gavves, E. 2022a. Nformer: Robust person re-identification with neighbor transformer. In *CVPR*, 7297–7307.

Wang, T.; Liu, H.; Song, P.; Guo, T.; and Shi, W. 2022b. Pose-guided Feature Disentangling for Occluded Person Re-identification Based on Transformer. In *AAAI*, volume 36, 2540–2549.

Wang, Z.; Huang, H.; Zheng, A.; Li, C.; and He, R. 2022c. Parallel Augmentation and Dual Enhancement for Occluded Person Re-identification. *arXiv preprint arXiv:2210.05438*.

Wang, Z.; Li, C.; Zheng, A.; He, R.; and Tang, J. 2022d. Interact, embed, and enlarge: Boosting modality-specific representations for multi-modal person re-identification. In *AAAI*, volume 36, 2633–2641.

Wang, Z.; Zhu, F.; Tang, S.; Zhao, R.; He, L.; and Song, J. 2022e. Feature Erasing and Diffusion Network for Occluded Person Re-Identification. In *CVPR*, 4754–4763.

Wu, A.; Zheng, W.-S.; Yu, H.-X.; Gong, S.; and Lai, J. 2017. RGB-infrared cross-modality person re-identification. In *ICCV*, 5380–5389.

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*, 4330–4339.

Zhang, G.; Zhang, Y.; Zhang, T.; Li, B.; and Pu, S. 2023. PHA: Patch-Wise High-Frequency Augmentation for Transformer-Based Person Re-Identification. In *CVPR*, 14133–14142.

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*, 7349–7358.

Zhang, Y.; and Wang, H. 2023. Diverse Embedding Expansion Network and Low-Light Cross-Modality Benchmark for Visible-Infrared Person Re-identification. In *CVPR*, 2153–2162.

Zheng, A.; Liu, J.; Wang, Z.; Huang, L.; Li, C.; and Yin, B. 2023. Visible-infrared person re-identification via specific and shared representations learning. *Visual Intelligence*, 1: 1–12.

Zheng, A.; Wang, Z.; Chen, Z.; Li, C.; and Tang, J. 2021. Robust Multi-Modality Person Re-identification. In *AAAI*, volume 35, 3529–3537.

Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; and Tian, Q. 2015. Scalable person re-identification: A benchmark. In *ICCV*, 1116–1124.

Zhou, K.; Yang, Y.; Cavallaro, A.; and Xiang, T. 2019. Omni-scale feature learning for person re-identification. In *ICCV*, 3702–3712.

Zhou, X.; Zhong, Y.; Cheng, Z.; Liang, F.; and Ma, L. 2023. Adaptive Sparse Pairwise Loss for Object Re-Identification. In *CVPR*, 19691–19701.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2223–2232.

Zhu, K.; Guo, H.; Zhang, S.; Wang, Y.; Huang, G.; Qiao, H.; Liu, J.; Wang, J.; and Tang, M. 2021. Aaformer: Auto-aligned transformer for person re-identification. *arXiv preprint arXiv:2104.00921*.