A Multi-Constraint Similarity Learning with Adaptive Weighting for Visible-Thermal Person Re-Identification

Yongguo Ling[1] , **Zhiming Luo**[1*] , **Yaojin Lin**[2] and **Shaozi Li**[1*]
[1]Department of Artificial Intelligence, Xiamen University, China
[2]School of Computer Science, Minnan Normal University, China
lingyongguo@stu.xmu.edu.cn, {zhiming.luo, szlig}@xmu.edu.cn, yjlin@mnnu.edu.cn

Abstract

The challenges of visible-thermal person re-identification (VT-ReID) lies in the inter-modality discrepancy and the intra-modality variations. An appropriate metric learning plays a crucial role in optimizing the feature similarity between the two modalities. However, most existing metric learning-based methods mainly constrain the similarity between individual instances or class centers, which are inadequate to explore the rich data relationships in the cross-modality data. Besides, most of these methods fail to consider the importance of different pairs, incurring an inefficiency and ineffectiveness of optimization. To address these issues, we propose a Multi-Constraint (MC) similarity learning method that jointly considers the cross-modality relationships from three different aspects, i.e., Instance-to-Instance (I2I), Center-to-Instance (C2I), and Center-to-Center (C2C). Moreover, we devise an Adaptive Weighting Loss (AWL) function to implement the MC efficiently. In the AWL, we first use an adaptive margin pair mining to select informative pairs and then adaptively adjust weights of mined pairs based on their similarity. Finally, the mined and weighted pairs are used for the metric learning. Extensive experiments on two benchmark datasets demonstrate the superior performance of the proposed over the state-of-the-art methods.

1 Introduction

Traditional person re-identification (ReID) [Zhong *et al.*, 2018; Zheng *et al.*, 2017; Yang *et al.*, 2020b] aims at matching a specific query person from large-scale gallery images captured by RGB cameras. However, the quality of the visible (RGB) image will significantly decrease under poor illumination (*e.g.* night-time). To overcome this issue, many thermal surveillance cameras have been deployed to capture thermal images. Hence, we encounter the task of matching person samples between visible and thermal cameras, which is known as visible thermal person re-identification (VT-ReID).

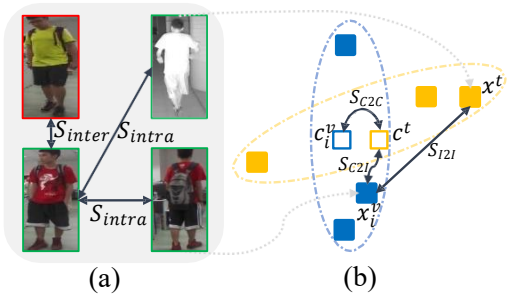

Figure 1: (a) Illustration of the key challenges in the VT-ReID task. The intra-class similarity S_{intra} often smaller than the inter-class similarity S_{inter}, due to the pose variations, inter-modality discrepancy. (b) Illustration of the result of ignoring the I2I constraint. The cross-modality intra-class similarity (S_{I2I}) maintains small when the model converged (The similarity between class centers S_{C2C} are optimized). Shapes indicate the identities and colors represent the modalities (blue for visible and yellow for thermal).

Apart from the long-standing inter-and intra-class variations suffered in the traditional REID, VT-ReID further needs to deal with the large inter-modality discrepancy (Figure 1(a)).

The key solution for VT-ReID is learning a shared embedding space in which the features of two modalities can be matched. Therefore, different metric-learning-based methods have been proposed to reduce the inter-modality discrepancy and intra-modality variations. These methods can be mainly divided into two aspects based on the similarity constraint utilized during the learning phase: (1) Instance-to-Instance (I2I): constraining the inter-and-intra modality similarity relationship of different training samples at the instance level [Feng *et al.*, 2019; Hao *et al.*, 2019c], and (2) Center-to-Center (C2C): constraining the relationship between the class centers from two different modalities [Liu and Tan, 2020; Zhu *et al.*, 2020]. Despite their success, the former I2I-based methods mainly focus on constraining the relationship between instances while ignoring the overall global class-level characters. They will be easily affected by noise or hard sample-pairs, resulting in inferior performance and generalization. On the other hand, the latter C2C-based methods indirectly optimize the feature similarity and significantly ignore intra-class variations of different samples. This may lead the situation that the similarity between class centers of

*Corresponding authors

two different modalities are optimized, but the discrepancy of cross-modality (S_{I2I}) still maintain large (Figure 1(b)). In a nutshell, these single constraint metric learning-based methods are insufficient to explore the rich data relationships among the large inter-modality discrepancy and intra-modality variations. Besides, most of these methods treat different informative pairs equally and fail to explore the cross-modality informative pairs for similarity learning, incurring an inefficiency and ineffectiveness of optimization.

To overcome the shortcomings of previous methods, we propose a Multi-Constraint (MC) similarity learning method to fully explore the various data relationship among cross-modality samples. In this study, we mainly consider the following three constraints, i.e., Center-to-Center (C2C), Center-to-Instance (C2I), and Instance-to-Instance (I2I). Specifically, the C2C constraint mainly enforces the cross-modality intra-class centers should have higher similarity than inter-class centers, whose primary goal is to bridge the large cross-modality gap. The C2I constraint further enforces that each sample should be close to its corresponding class center of the other modality to reduce the intra-modality variations. Additionally, the I2I constraint force the inter-modality intra-class discrepancy between samples should be smaller by pulling them close to each other. These three constraints are jointly utilized to reduce the intra-class discrepancy and enlarge the inter-class distance.

In addition, different informative pairs have different effects during similarity learning, and we devise an Adaptive Weighting Loss (AWL) function to effectively implement the multi-constraint. In the AWL, we first use an adaptive margin to select cross-modality informative pairs and discard those less informative pairs. Then, we exploit a sigmoid variant function to adjust pair weights based on their similarity. By utilizing this pair mining and pair weighting scheme, we will encourage the network to pay more attention to optimize the informative pairs during training.

To sum up, our main contributions are as follows: 1) We propose a Multi-Constraint (MC) similarity learning framework to reduce the inter-and intra-modality discrepancy in the VT-ReID by jointly considering three constraints (I2I, C2I, and C2C). 2) We devise an Adaptive Weighting Loss (AWL) function to leverage the cross-modality informative pairs for model training in a more effective manner. 3) Experiments on two datasets demonstrate the mutual benefits of the above-proposed components and show superior performance over state-of-the-art methods.

2 Related Work

The existing methods of VT-ReID can be mainly divided into three groups. **1) Feature extraction based methods.** [Wu *et al.*, 2017] first introduce SYSU-MM01, a visible thermal dataset, and they study a deep zero-padding method for evolving domain-specific nodes in the network. [Ye *et al.*, 2018b] introduce a two streams network to learn cross-modality embedding. To handle the cross-modality discrepancy, [Lu *et al.*, 2020] introduce a cross-modality shared-specific feature network, [Ye *et al.*, 2020] develop a dynamic dual-attentive aggregation network, and [Yang *et al.*, 2020a] propose a bi-directional random walk scheme network. These methods mainly focus on exploring a cross-modality network to extract discriminative feature. **2) Metric learning based methods** [Dai *et al.*, 2018; Ye *et al.*, 2018a; Hao *et al.*, 2019c; Feng *et al.*, 2019; Hao *et al.*, 2019a; Hao *et al.*, 2019b; Ling *et al.*, 2020; Zhu *et al.*, 2020; Liu and Tan, 2020] are proposed to learn an embedding space that makes the intra-class samples close to each other. [Hao *et al.*, 2019c] and [Feng *et al.*, 2019] only consider the relationship between instances. Hence, [Zhu *et al.*, 2020] and [Liu and Tan, 2020] are proposed to optimize the relationship between class centers of two modalities. However, these methods with a single relationship constraint are not enough to express the rich data relationships. Therefore, we propose a Multi-Constraint (MC), which adopts the relationships of I2I, C2I, and C2C constraints to optimize the feature. **3) Image generation based methods.** [Wang *et al.*, 2019c; Wang *et al.*, 2019a; Li *et al.*, 2020] use a variational autoencoder and generative adversarial network to generate fake images to bridge the modality gap. These methods may contain id-unrelated factors, which undermines the performance, so [Choi *et al.*, 2020], [Wang *et al.*, 2020], and [Pu *et al.*, 2020] combine a disentangle representation with VAE for robust cross-modality matching. These models are commonly difficult to train due to complex generative adversarial networks to generate fake images.

3 Proposed Method

In this section, we will introduce the details of the proposed framework. As shown in Figure 2, the proposed framework mainly includes the base feature extractor, part-based embeddings, and training loss function. The training loss function consists of a multi-constraint based metric learning loss function L_{MC} and the identification loss function L_{ID}.

3.1 Base Feature Extractor and Part-Based Embeddings

Due to the large modality discrepancy, we first use two independent modules to compute the modality-specific features at the shallow layers. Then, we use parameter sharing module to learn the modality-shareable feature to embed the two modalities into the same subspace. The independent modules have the same structure as the first convolution block and residual block 1 (Layer1) in ResNet-50 [He *et al.*, 2016], and the share module is the same as the residual block 2-4 (Layer2-4).

Besides, the PCB method [Sun *et al.*, 2018] has demonstrated the effectiveness of using local parts to increase feature discrimination. Therefore, we exploit an efficient module to extract two parallel part features at different scales. Firstly, we divide the feature maps into six non-overlapping parts and three non-overlapping parts along the vertical direction. Then, we compute the feature of each local part by a global pooling and a convolution. The final local feature dimension of the branch with six parts and three parts are 256 and 512, respectively. Finally, the holistic features for these two-part branches are simply concatenating the corresponding local part features. In the testing stage, we concatenate these two holistic features ($[0.6f_{h3}, 0.4f_{h6}]$) to match the visible and thermal images.

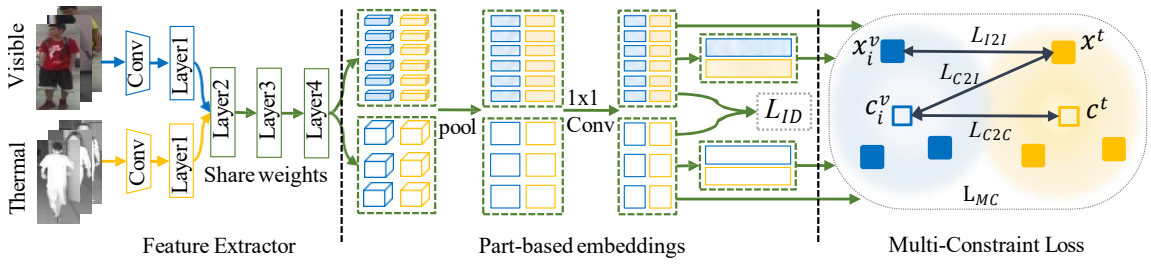

Figure 2: The framework of our proposed method includes feature extractor, multi-scale part, and Multi-Relationship Constraints. The feature extractor contains two independent modules and a sharing module. We use feature extractor to extract multi-scale part feature, which includes six-part features, three-part features, and their corresponding holistic features. Then, we optimize the network with cross-entropy identity loss (L_{ID}) and Multi-Constraint (MC). The MC adopts the relationships of Instance-to-Instance (I2I), Center-to-Instance (C2I), and Center-to-Center (C2C) constraints to jointly reduce the cross-modality discrepancy and increase the similarity of the same class. Shapes indicate the identities and colors represent the modalities (blue for visible and yellow for thermal).

3.2 The Multi-Constraint Loss

For learning a more discriminative feature, we propose a general metric learning by considering multiple cross-modality constraints, i.e., Center-to-Center (C2C), Center-to-Instance (C2I), and Instance-to-Instance (I2I). Specifically, the I2I and C2I constraints are leveraged to optimize the holistic features, the C2C constraint is utilized to constrain the local part features. These three constraints jointly reduce the inter-modality discrepancy and intra-class variations, and an illustration of them are shown in Figure 3.

Assume x_i^v and x_i^t are a sample of class i from visible modality and thermal modality, respectively. Their class center c_i^v and c_i^t can be computed by:

$$c_i^v = \frac{1}{C_i}\sum_{j=1}^{C_i} f_j^v, \quad c_i^t = \frac{1}{C_i}\sum_{j=1}^{C_i} f_j^t, \tag{1}$$

where C_i is the number of samples of class i for a modality in a mini-batch. f_j^v and f_j^t are the feature of x_j^v and x_j^v.

1) C2C Constraint. For handling the large inter-modality discrepancy, we use C2C constraint enforces the cross-modality intra-class centers have higher similarity than inter-class centers. Given a class center anchor of one modality (i.e., c_a^v), we select a positive and negative class center form another modality (c_p^t and c_n^t). Then, we reduce the distance between c_a^v and c_p^t, and enlarge the distance between c_a^v and c_n^t (Figure 3(a)). Given the selected positive and negative class center pairs, the loss of C2C constraint is defined as:

$$L_{C2C} = L(P_{C2C}, N_{C2C}), \tag{2}$$

where the P_{C2C} and N_{C2C} are positive and negative pairs satisfying C2C constraint, respectively. The pair-based loss function $L(P, N)$ will be introduced later.

2) C2I Constraint. We further enforces that each sample should be close to its corresponding class center of the other modality to reduce the intra-modality variations. Given a class center anchor of one modality (i.e., c_a^v), we select a positive and negative samples form another modality (x_p^t and x_n^t). Then, we force positive samples x_p^t to approach their own inter-modality class center c_a^v, and enlarge the distance between c_a^v and x_n^t (Figure 3(b)). The loss of C2I constraint is defined as follows:

$$L_{C2I} = L(P_{C2I}, N_{C2I}), \tag{3}$$

where the P_{C2I} and N_{C2I} indicate the positive and negative pairs of the C2I constraint, respectively.

3) I2I Constraint. We utilize the I2I constraint to increase the similarity of two matching samples from the two different modalities. Given a anchor sample of one modality (i.e., x_a^v), we select a positive and negative samples from another modality (x_p^t and x_n^t). Then, we pull the positive pairs close to each other and push the negative pairs away from each other. The loss of I2I constraint is defined as follows:

$$L_{I2I} = L(P_{I2I}, N_{I2I}), \tag{4}$$

where the P_{I2I} and N_{I2I} indicate the positive and negative pairs of the I2I constraint, respectively.

By jonitly considering these three constraints, the equation of MC is defined as follows:

$$L_{MC} = \alpha L_{I2I} + \beta L_{C2I} + \omega L_{C2C}, \tag{5}$$

where α, β, and ω are hyper-parameters.

3.3 Adaptive Weighting Loss

For implementing the $L(P, N)$ used in the multi-constraints, we devise a new adaptive weighting loss based on the Multi-Similarity (MS) Loss [Wang *et al.*, 2019b] for the cross-modality VT-ReID task.

Revisit Multi-Similarity (MS) Loss

The MS loss optimizes the feature space based on informative pairs mined based on three similarities: Self-similarity, Negative relative similarity, and Positive relative similarity. In the MS loss, And they propose a MS loss to optimize sample pairs by considering these three similarities. Set x_i as an anchor, a negative and positive pair $\{x_i, x_j\}$ are selected if the pair similarity S_{ij} satisfies the following conditions:

$$S_{ij}^- > \min_{y_k=y_i} S_{ik} - m, S_{ij}^+ < \max_{y_k\neq y_i} S_{ik} + m \tag{6}$$

where m is a given margin.

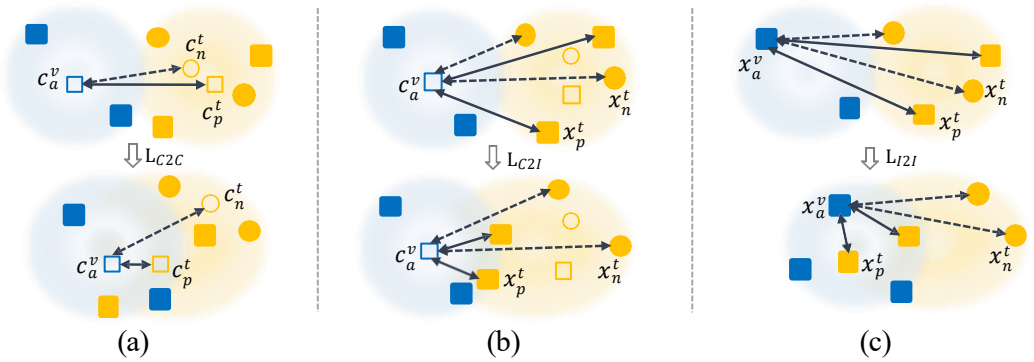

Figure 3: Illustration of the three constraints of proposed MC. (a) Given a anchor class center of one modality (c_a^v), we select a positive and negative class center form another modality (c_p^t and c_n^t). Then, we reduce the distance between c_a^v and c_p^t, and enlarge the distance between c_a^v and c_n^t. (b) We also select a positive and negative samples form another modality (x_p^t and x_n^t). Then, we increase the positive pairs similarity, and reduce the negative pairs similarity. (c) Given a anchor sample of one modality (x_a^v), we select a positive and negative samples form another modality (x_p^t and x_n^t). Then, we pull the positive pairs close to each other as well as push the negative pairs away from each other.

The MS loss function can be calculated by:

$$\mathcal{L}_{\mathcal{MS}} = \frac{1}{C}\sum_{i=1}^{C}\left\{\frac{1}{\lambda_p}\log\left[1+\sum_{k\in\mathcal{P}_i} e^{-\lambda_p(S_{ik}-\lambda)}\right] + \frac{1}{\lambda_n}\log\left[1+\sum_{k\in\mathcal{N}_i} e^{\lambda_n(S_{ik}-\lambda)}\right]\right\}, \tag{7}$$

where C is the number of selected anchors. λ_p, λ_n, and λ are hyper-parameters.

Despite the effectiveness of MS loss in the standard single modality task, the MS will suffer the issue of discarding many informative pairs for the cross-modality VT-ReID. As shown in figure 4 (a), the negative pair with maximum cross-modality similarity is the a^v, n^t. The similarity of positive pair a^v, p^t is larger than the a^v, n^t which does not satisfy with the condition in Eq. 6. Then, this informative positive pair will be discarded and not be used for optimization. To deal with this issue, we propose a new pair mining strategy for selecting cross-modality informative pairs. Besides, we further devise a weighting function to adjust the weights of selected pairs.

Pair Mining

For a given visible modality anchor a^v, the p^t and n^v are a positive sample and a negative sample from the inter-modality, respectively. Then we will decide whether selecting p^t or n^v to form a positive pair or a negative pair of a^v based on following conditions.

For the selection of positive pair, we first find the closest negative intra-modality sample n^v of p^t. If the similarity of $\{n^v, p^t\}$ is higher than the $\{a^v, p^t\}$, it means that the similarity of $\{a^v, p^t\}$ is not optimized as shown in Fig.4. We will select $\{a^v, p^t\}$ as a positive pair. Formally, the condition for the positive pair selection with an addition margin can be denote as,

$$S\left(a^v, p^t\right) < \max_{n^v} S\left(p^t, n^v\right) + m, \tag{8}$$

where m is a margin.

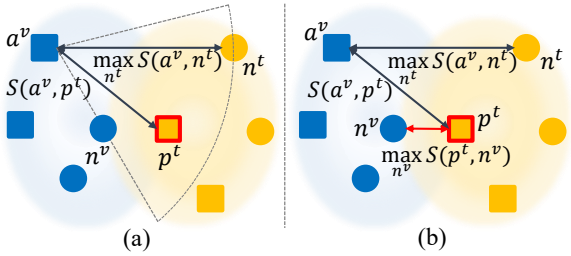

Figure 4: Illustration of the pair mining of MS loss and our AWL. (a) Set a^v as the anchor, the inter-modality hard positive samples p^t (red borer) are not selected following the mining of MS loss. (b) The informative pair (red borer) can be selected following our pair mining condition.

In a similar manner, a inter-modality negative pair $\{a^v, n^t\}$ is select if $S(a^v, n^t)$ satisfies following conditions:

$$S\left(a^v, n^t\right) > \min_{n^v} S\left(n^t, n^v\right) - m, \tag{9}$$

where the n^v and n^t are belong the same label.

Pair Weighting

After selecting the informative positive and negative pairs, we will adaptively assign large weights for more informative pairs, impelling the model to pay more attention to the informative pair during training. In this step, we exploit a sigmoid variant function to adaptively adjusting weighting based on their similarity with the anchor. Specifically, the weights are inversely correlated to positive pairs similarity and positively correlated to negative pairs similarity. The adaptive weighting function for positive pairs and negative pairs are defined as follows:

$$W_p = 10 \times \delta(-\tfrac{S(a,p)-0.5}{0.5}), \quad W_n = 3 \times \delta(\tfrac{S(a,n)-0.5}{0.5}), \tag{10}$$

where δ indicates the sigmoid function.

We then design the Adaptive Weighting Loss (AWL) based on the Eq 7, by considering the selected informative pairs and

Methods	Venue	All-search			
		Single-shot		Multi-shot	
		R1	mAP	R1	mAP
D2RL	CVPR19	28.9	29.2	/	/
DGD	TIP19	37.35	38.11	43.86	30.48
AlignGAN	ICCV19	42.4	40.7	51.5	33.9
DFE	MM19	48.71	48.59	54.63	42.14
Hi-CMD	CVPR20	34.94	35.94	/	/
PIG	AAAI20	38.1	36.9	45.1	29.5
cmSSFT	CVPR20	47.7	54.1	57.4	59.1
Xmodal	AAAI20	49.92	50.73	/	/
CML	MM20	51.8	51.21	56.27	43.39
DDAG	ECCV20	54.75	53.02	/	/
DG-VAE	MM20	59.49	58.46	/	/
SIM	CVPR20	60.88	56.93	/	/
HC-TRI	TMM21	61.68	57.51	/	/
Ours	IJCAI21	**64.82**	**60.81**	**68.05**	**51.48**

Table 1: Comparison with the state-of-the-art methods on the SYSU-MM01 dataset.

their adaptive weights. The final AWL is defined as:

$$L(\mathcal{P},\mathcal{N}) = \frac{1}{C}\sum_{i=1}^{C}\left\{\log\left[1+\sum_{k\in\mathcal{P}_i} e^{-W_p(S(a,p)-\lambda)}\right] + \log\left[1+\sum_{z\in\mathcal{N}_i} e^{W_n(S(a,n)-\lambda)}\right]\right\}, \tag{11}$$

where C is the number of selected anchors. λ is a hyper-parameter.

3.4 Final Loss Function

Finally, we train our proposed framework by combining the Multi-Constraint (MC) loss L_{MC} and the identity loss L_{ID} in an end-to-end manner. The loss is as follows:

$$\arg\min_{\theta} L_{MC} + \gamma L_{ID}, \tag{12}$$

where θ represents the parameters of the model, and δ is a hyper-parameter controlling the influence of the two losses.

4 Experiments

4.1 Experimental Settings

Datasets. We evaluate our proposed methods on two publicly available VT-ReID datasets (SYSU-M001 [Wu *et al.*, 2017] and RegDB [Nguyen *et al.*, 2017]). **SYSU-M001** contains 287,628 RGB images and 15,729 infrared images captured by four RGB cameras and two thermal cameras. The training set consists of 22,258 RGB images and 11,909 infrared images from 395 identities. **RegDB** contains 4,120 RGB images and 4,120 infrared images. There are 412 identities, where 206 identities for training and others for testing.

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

Setting		V2T		T2V	
Method	Venue	R1	mAP	R1	mAP
D2RL	CVPR19	/	/	43.4	44.1
AlignGAN	ICCV19	57.9	53.6	56.3	53.4
DFE	MM19	70.13	69.14	67.99	66.70
Hi-CMD	CVPR20	/	/	70.93	66.04
PIG	AAAI20	48.50	49.3	48.1	48.90
Xmodal	AAAI20	62.21	60.18	/	/
CML	MM20	59.81	60.86	/	/
cmSSFT	CVPR20	65.4	65.6	63.8	64.2
DDAG	ECCV20	69.34	63.46	68.06	61.80
DG-VAE	MM20	72.97	71.78	/	/
SIM	CVPR20	75.29	74.47	78.30	75.24
HC-TRI	TMM21	91.05	83.28	89.30	81.46
Ours	IJCAI21	**93.83**	**87.55**	**91.55**	**85.25**

Table 2: Comparison with the state-of-the-art methods on the RegDB dataset on visible-thermal and thermal-visible settings.

Method	RegDB		SYSU-M001	
	R1	mAP	R1	mAP
Baseline(w/ Part)	64.71	63.10	48.15	48.39
+MC	91.55	85.25	64.82	60.81
Baseline(w/o Part)	40.34	38.16	35.55	34.97
+MC	54.81	54.12	50.25	48.63

Table 3: Evaluation of the MC under the baseline with part (w/Part) and without part (wo/Part).

Implementation details. The baseline network is constructed based on [Liu and Tan, 2020], which optimizes part features with the identity loss and hc-tri loss. We also evaluate the performance of a baseline framework without part ([Ye *et al.*, 2018b]) to demonstrate the generalization of our proposed. The training batch is set to 96 (48 RGB images and 48 infrared images from 6 person IDs) and 64 (32 RGB images and 32 infrared images from 8 person IDs) for SYSU-MM01 and RegDB dataset, respectively. The input images are resized to $288 \times 144 \times 3$ for both RGB and infrared images. We use the SGD optimizer for training with an initial learning rate of 0.01, and train the model for 80 epochs. We divide the learning rate by 10 after every 10 epochs. The weights α, β, and ω in Eq. 5 are set to 0.5, 1.0, and 0.2, respectively. The pair mining margin m in Eq. 8 is set to 0.2. λ in Eq. 11 is set to 0.5. The weights γ in Eq. 12 is set to 1.

Method	RegDB		SYSU-M001	
	R1	mAP	R1	mAP
Baseline(w/o Part)	40.34	38.16	35.55	34.97
+I2I	50.63	51.72	47.44	47.79
+C2I	51.70	51.99	48.59	46.92
+C2C	51.02	50.50	47.25	47.41
+I2I+C2C	51.26	52.22	49.01	47.67
+C2I+C2C	52.33	52.25	49.49	48.06
+I2I+C2I	53.30	52.23	49.41	49.81
+I2I+C2I+C2C	54.81	54.12	50.25	48.63

Table 4: Investigation of three constraints in MC.

Method	RegDB		SYSU-M001	
	R1	mAP	R1	mAP
Baseline(w/o Part)	40.34	38.16	35.55	34.97
+MS (MS mining)	44.85	41.74	40.07	39.29
+MS (Our mining)	49.32	48.60	44.12	45.35
+AWL (MS mining)	46.75	46.03	41.65	40.39
+AWL (w/o weighting)	49.85	49.04	44.46	46.39
+AWL	50.63	51.72	47.44	47.79

Table 5: Evaluation of the pair mining and pair weighting in our proposed AWL with I2I constraint.

Method	RegDB		SYSU-M001	
	R1	mAP	R1	mAP
Baseline(w/o Part)	40.34	38.16	35.55	34.97
+Triplet Loss(Hard)	42.28	42.49	42.02	42.84
+MS Loss(cm-Mining)	45.29	41.96	41.55	43.26
+Hc-Tri Loss	48.59	45.22	38.84	38.90
+CML Loss	50.29	48.34	44.36	44.25
+Our Loss	54.81	54.12	50.25	48.63

Table 6: Comparison of the MC with other metric-learning methods.

4.2 Comparison with State of The Art

We compare our proposed method with recently published state of the arts on SYSU-MM01 and RegDB datasets. The comparison includes: feature extraction based methods (cmSSFT with single query [Lu *et al.*, 2020], DDAG [Ye *et al.*, 2020], SIM [Jia *et al.*, 2020]), metric learning based methods (DGD [Feng *et al.*, 2019], DFE [Hao *et al.*, 2019a], CML [Ling *et al.*, 2020], HC-TRI [Liu and Tan, 2020]), image generation based methods (D2RL [Wang *et al.*, 2019c], AlignGAN [Wang *et al.*, 2019a], Xmodal [Li *et al.*, 2020], Hi-CMD [Choi *et al.*, 2020], PIG [Wang *et al.*, 2020], DG-VAE [Pu *et al.*, 2020]).

The results on SYSU-MM001 and RegDB datasets are shown in Tabel 1 and Tabel 2, respectively. We can see that proposed method outperforms the state-of-the-art methods on both datasets. Specifically, our method achieve **rank-1 accuracy = 64.82** and **mAP accuracy = 60.81** on SYSU-MM01 for the single-shot setting of the all-search mode, and, **rank-1 accuracy = 93.83** and **mAP accuracy = 87.55** on RegDB for visible to thermal setting.

4.3 Evaluation

In this section, we evaluate the effectiveness of each component in our proposed method on SYSU-MM01 (single shot setting of all-search mode) and RegDB (thermal to visible setting).

Comparison over the baseline. To evaluate the effectiveness and demonstrate the generalization of our proposed Multi-Constraints (MC), we conduct experiments of ablation study under two settings: the baseline with part (w/ Part) [Liu and Tan, 2020] and without part (w/o Part) [Ye *et al.*, 2018b] on both datasets. The results show in Table 3. We can see that the proposed MC can significantly improve the performance over the baseline on both datasets for two settings. The observation verifies the effectiveness and generalization of our proposed MC for the challenge cross-modality VT-ReID.

Investigation of three constraints in MC. In Table 4, we further evaluate the effectiveness of different constraints in the MC. From the table, we can find that considering every single constraint can significantly improve the performance over the baseline. Moreover, combining two of them can further improve the performance. Additionally, the highest accuracy is obtained by jointly considering the three constraints. These results demonstrate the complementary and mutual benefits of the three constraints in MC.

Evaluation of the components in AWL. To evaluate the effectiveness of the proposed Adaptive Weighting Loss (AWL), we conduct a controlled experiment with MS loss [Wang *et al.*, 2019b]. The results are reported in Table 5. By replacing the pair mining in MS loss with ours (+MS (our mining) vs. +MS (MS mining)) and the pair mining in our AWL with the one in MS (+AWL (MS mining) vs. +AWL), we can observe a significant performance boost for the former case and a notable degeneration for the later. Then, by comparing the AWL without pair weighting to AWL (+AWL (w/o weighting) vs. +AWL), we can find that AWL with pair weighting gain a higher result than AWL without pair weighting. These results demonstrate the effectiveness of the proposed pair mining and pair weighting strategies.

Comparison of the MC with other metric-learning methods. In order to demonstrate the superiority of our proposed MC, we further compare our MC with other various different metric-learning loss functions, including Triplet Loss (Hard) [Hermans *et al.*, 2017] (I2I), MS Loss with cross-modality pair mining (cm-Mining) [Wang *et al.*, 2019b] (I2I), Hc-Tri Loss [Liu and Tan, 2020] (C2C), and CML Loss [Ling *et al.*, 2020] (C2C & C2I). Notice that these loss functions usually consider one or two constraints. By comparing our MC with them, our method outperforms them by a large margin. These results suggest considering more comprehensive data relationships is essential for overcoming the large inter-modality discrepancy and intra-and-inter class variations in VT-ReID.

5 Conclusion

In this paper, we propose a Multi-Constraint (MC) for VT-ReID, which jointly considers Instance-to-Instance (I2I), Center-to-Instance (C2I), and Center-to-Center constraints (C2C) to optimize feature similarity to reduce inter-modality discrepancy and intra-modality variations. Furthermore, we implement MC with an Adaptive Weighting Loss (AWL) to promote the model training. Extensive experiments on two VT-ReID datasets demonstrate the superior performance of the proposed over the state of the art methods.

Acknowledgements

This work is supported by the National Nature Science Foundation of China (No. 61876159, No. 61806172, No. 62076116, No. U1705286), the China Postdoctoral Science Foundation Grant (No. 2019M652257).

References

[Choi *et al.*, 2020] Seokeon Choi, Sumin Lee, Youngeun Kim, Taekyung Kim, and Changick Kim. Hi-cmd: Hierarchical cross-modality disentanglement for visible-infrared person re-identification. In *CVPR*, 2020.

[Dai *et al.*, 2018] Pingyang Dai, Rongrong Ji, Haibin Wang, Qiong Wu, and Yuyu Huang. Cross-modality person re-identification with generative adversarial training. In *IJCAI*, 2018.

[Feng *et al.*, 2019] Zhanxiang Feng, Jianhuang Lai, and Xiaohua Xie. Learning modality-specific representations for visible-infrared person re-identification. *TIP*, 2019.

[Hao *et al.*, 2019a] Yi Hao, Nannan Wang, Xinbo Gao, Jie Li, and Xiaoyu Wang. Dual-alignment feature embedding for cross-modality person re-identification. In *ACM MM*, 2019.

[Hao *et al.*, 2019b] Yi Hao, Nannan Wang, Xinbo Gao, Jie Li, and Xiaoyu Wang. Dual-alignment feature embedding for cross-modality person re-identification. In *ACM MM*, 2019.

[Hao *et al.*, 2019c] Yi Hao, Nannan Wang, Jie Li, and Xinbo Gao. Hsme: hypersphere manifold embedding for visible thermal person re-identification. In *AAAI*, 2019.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[Hermans *et al.*, 2017] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[Jia *et al.*, 2020] Mengxi Jia, Yunpeng Zhai, Shijian Lu, Siwei Ma, and Jian Zhang. A similarity inference metric for rgb-infrared cross-modality person re-identification. In *IJCAI*, 2020.

[Li *et al.*, 2020] Diangang Li, Xing Wei, Xiaopeng Hong, and Yihong Gong. Infrared-visible cross-modal person re-identification with an x modality. In *AAAI*, 2020.

[Ling *et al.*, 2020] Yongguo Ling, Zhun Zhong, Zhiming Luo, Paolo Rota, Shaozi Li, and Nicu Sebe. Class-aware modality mix and center-guided metric learning for visible-thermal person re-identification. In *ACM MM*, 2020.

[Liu and Tan, 2020] Haijun Liu and Xiaoheng Tan. Parameters sharing exploration and hetero-center based triplet loss for visible-thermal person re-identification. *arXiv preprint arXiv:2008.06223*, 2020.

[Lu *et al.*, 2020] Yan Lu, Yue Wu, Bin Liu, Tianzhu Zhang, Baopu Li, Qi Chu, and Nenghai Yu. Cross-modality person re-identification with shared-specific feature transfer. In *CVPR*, 2020.

[Nguyen *et al.*, 2017] Dat Nguyen, Hyung Hong, Ki Kim, and Kang Park. Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors*, 2017.

[Pu *et al.*, 2020] Nan Pu, Wei Chen, Yu Liu, Erwin M Bakker, and Michael S Lew. Dual gaussian-based variational subspace disentanglement for visible-infrared person re-identification. In *ACM MM*, 2020.

[Sun *et al.*, 2018] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, 2018.

[Wang *et al.*, 2019a] Guan'an Wang, Tianzhu Zhang, Jian Cheng, Si Liu, Yang Yang, and Zengguang Hou. Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment. In *ICCV*, 2019.

[Wang *et al.*, 2019b] Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R Scott. Multi-similarity loss with general pair weighting for deep metric learning. In *CVPR*, 2019.

[Wang *et al.*, 2019c] Zhixiang Wang, Zheng Wang, Yinqiang Zheng, Yung-Yu Chuang, and Shin'ichi Satoh. Learning to reduce dual-level discrepancy for infrared-visible person re-identification. In *CVPR*, 2019.

[Wang *et al.*, 2020] Guan-An Wang, Tianzhu Zhang Yang, Jian Cheng, Jianlong Chang, Xu Liang, Zengguang Hou, et al. Cross-modality paired-images generation for rgb-infrared person re-identification. In *AAAI*, 2020.

[Wu *et al.*, 2017] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. Rgb-infrared cross-modality person re-identification. In *ICCV*, 2017.

[Yang *et al.*, 2020a] Fan Yang, Zheng Wang, Jing Xiao, and Shin'ichi Satoh. Mining on heterogeneous manifolds for zero-shot cross-modal image retrieval. In *AAAI*, 2020.

[Yang *et al.*, 2020b] Fengxiang Yang, Ke Li, Zhun Zhong, Zhiming Luo, Xing Sun, Hao Cheng, Xiaowei Guo, Feiyue Huang, Rongrong Ji, and Shaozi Li. Asymmetric co-teaching for unsupervised cross-domain person re-identification. In *AAAI*, 2020.

[Ye *et al.*, 2018a] Mang Ye, Xiangyuan Lan, Jiawei Li, and Pong C Yuen. Hierarchical discriminative learning for visible thermal person re-identification. In *AAAI*, 2018.

[Ye *et al.*, 2018b] Mang Ye, Zheng Wang, Xiangyuan Lan, and Pong C Yuen. Visible thermal person re-identification via dual-constrained top-ranking. In *IJCAI*, 2018.

[Ye *et al.*, 2020] Mang Ye, Jianbing Shen, David J Crandall, Ling Shao, and Jiebo Luo. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In *ECCV*, 2020.

[Zheng *et al.*, 2017] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, 2017.

[Zhong *et al.*, 2018] Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang. Camera style adaptation for person re-identification. In *CVPR*, 2018.

[Zhu *et al.*, 2020] Yuanxin Zhu, Zhao Yang, Li Wang, Sai Zhao, Xiao Hu, and Dapeng Tao. Hetero-center loss for cross-modality person re-identification. *Neurocomputing*, 2020.