A Similarity Inference Metric for RGB-Infrared Cross-Modality Person Re-identification

Mengxi Jia[1] , **Yunpeng Zhai**[1] , **Shijian Lu**[2] , **Siwei Ma**[3,4] and **Jian Zhang**[1,4*]
[1]School of Electronic and Computer Engineering, Peking University, China
[2]Nanyang Technological University, Singapore
[3]School of Electronics Engineering and Computer Science, Peking University, China
[4] Peng Cheng Laboratory, China
{mxjia, ypzhai, swma, zhangjian.sz}@pku.edu.cn, shijian.lu@ntu.edu.sg

Abstract

RGB-Infrared (IR) cross-modality person re-identification (re-ID), which aims to search an IR image in RGB gallery or vice versa, is a challenging task due to the large discrepancy between IR and RGB modalities. Existing methods address this challenge typically by aligning feature distributions or image styles across modalities, whereas the very useful similarities among gallery samples of the same modality (i.e. intra-modality sample similarities) are largely neglected. This paper presents a novel similarity inference metric (SIM) that exploits the intra-modality sample similarities to circumvent the cross-modality discrepancy targeting optimal cross-modality image matching. SIM works by successive similarity graph reasoning and mutual nearest-neighbor reasoning that mine cross-modality sample similarities by leveraging intra-modality sample similarities from two different perspectives. Extensive experiments over two cross-modality re-ID datasets (SYSU-MM01 and RegDB) show that SIM achieves significant accuracy improvement but with little extra training as compared with the state-of-the-art.

1 Introduction

Person re-identification (re-ID) is an important task in video surveillance. Given a query image of a person, re-ID aims to match persons in an image gallery collected from non-overlapping camera networks.

To leverage the unique features of sensors of different modalities, cross-modality re-ID has been attracting increasing interest in recent years for more robust identification, e.g. by using infrared images as query and RGB images as gallery. On the other hand, cross-modality re-ID remains a challenging task due to the large discrepancy across image modalities in terms of distinct illumination, heterogeneous features, etc.

Two typical approaches have been explored to address the cross-modality re-ID challenges. The first approach attempts to align feature distribution of images of different modalities to reduce the cross-modality discrepancy [Wu *et al.*, 2017] [Ye *et al.*, 2018a] [Ye *et al.*, 2018b] [Dai *et al.*, 2018]. The other approach utilizes generative adversarial network (GAN) as a modality transformer to convert person images from one modality to another while preserving the identity information as much as possible [Wang *et al.*, 2019b][Wang *et al.*, 2020][Wang *et al.*, 2019a]. These two types of approaches thus focus more on the reduction of cross-modality discrepancy or learning ID-preservative mapping across modalities, whereas the discriminative similarity among gallery samples of the same modality is largely neglected. The lack of this very useful information has become a major reason for the low performance of cross-modality person re-ID.

In this paper, we propose an innovative similarity inference metric (SIM) for cross-modality person re-ID. SIM aims to infer cross-modality sample similarities by exploiting reliable intra-modality sample similarities as illustrated in Fig. 1. Instead of using the query-gallery similarity matrix for person matching like most existing methods do, we introduce sim-

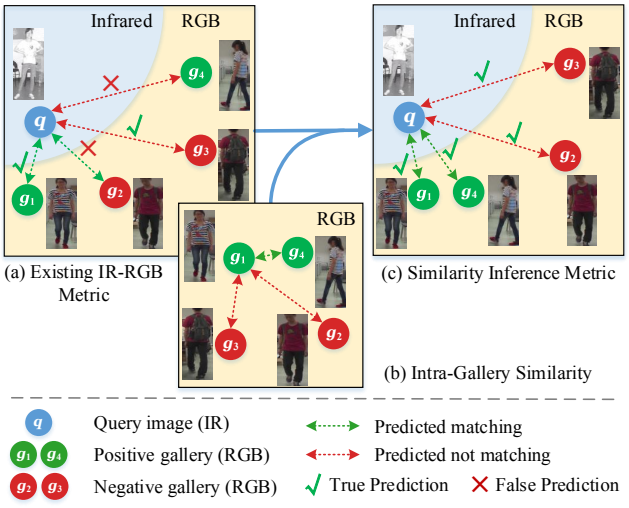

Figure 1: Similarity Inference Metric (SIM) infers cross-modality sample similarities by exploiting intra-modality sample similarities. It enhances the existing IR-RGB metric to match the hard positive samples (*e.g.* g_4) which are dissimilar from the query but similar to the predicted matching samples of the query (*e.g.* g_1).

*Corresponding author.

ilarity graph reasoning (SGR) and mutual nearest-neighbor reasoning (MNNR) that discover intra-modality sample similarities and circumvent the cross-modality discrepancy successively. Specifically, these two types of reasoning utilize the intra-modality similarities, in terms of graph shortest path and nearest neighbor overlap, to empower re-ID to match the hard positive samples which are dissimilar from the query but similar to the predicted matching samples of the query. What's more, SIM improves the cross-modality re-ID performance significantly and consistently. More details will be provided in the experimental section.

The main contributions of this work can be summarized in three aspects. *First*, it proposes a similarity inference metric that successively improves cross-modality similarities by utilizing the discriminative intra-modality sample similarities. *Second*, it designs novel similarity graph reasoning and mutual nearest-neighbor reasoning that can be applied to different cross-modality person re-ID metric with little extra training. *Third*, it achieves significant performance improvement over the state-of-the-arts on two widely used cross-modality re-ID datasets: SYSU-MM01 and RegDB.

2 Related Works

2.1 Single-modality Person Re-ID

Conventional re-ID research mainly focuses on the challenge of appearance variations in a single RGB modality, including illumination conditions, viewpoint variations, misalignment, etc. Existing methods can be broadly classified into two categories. Methods in the first category attempt to learn similarity metrics which are used to predict whether two images contain the same person [Zheng *et al.*, 2011] [Zhen *et al.*, 2013] [Gou *et al.*, 2014] [Liao *et al.*, 2015] [Chen *et al.*, 2017][Hermans *et al.*, 2017]. Methods in the second category focus on learning a discriminative feature representation, upon which efficient L2 or cosine distances can be applied [Liao *et al.*, 2015] [Zhao *et al.*, 2017] [Li *et al.*, 2018] [Zhai *et al.*, 2020] [Yang *et al.*, 2019]. Besides, most existing methods were developed for single-modality re-ID which cannot tackle the cross-modality re-ID well due to the large discrepancy across modalities.

2.2 Cross-modality Person Re-ID

For the RGB-Infrared cross-modality re-ID, the discrepancies come not just from appearance variations but also from heterogeneous images of different modalities. Two typical approaches have been explored to reduce the cross-modality discrepancies. The first approach attempts to align the feature distribution of images of different modalities. For example, [Wu *et al.*, 2017] explores three different network structures and uses deep zero-padding for evolving domain-specific nodes. [Ye *et al.*, 2018a] jointly optimizes the modality-specific and modality-shared metrics to learn multi-modality representations. [Ye *et al.*, 2018b] proposes a dual-path network with a bi-directional dual-constrained top-ranking loss to learn common representations. [Dai *et al.*, 2018] designs a cross-modality generative adversarial network (cmGAN) to learn discriminative representations from different modalities. [Hao *et al.*, 2019] proposes a hyper-sphere manifold embedding model. The second approach instead uses generative adversarial network (GAN) as a modality transformer to convert person images from one modality to another while preserving the identity information as much as possible [Wang *et al.*, 2019b] [Wang *et al.*, 2019a] [Wang *et al.*, 2020].

Though these methods reduce the modality discrepancies, the very useful discriminative similarity among gallery samples of the same modality is largely neglected. Our similarity inference metric captures such intra-gallery sample similarity which improves the cross-modality re-ID significantly, more details to be discussed in the ensuing subsections.

2.3 Re-ranking for Person Re-ID

Re-ranking methods have been wildly studied to improve conventional person re-ID. After an initial ranking list is obtained, re-ranking aims to raise the rank of relevant images in an automatic and unsupervised manner. Recently, various re-ranking methods have been reported. For example, [Garcia *et al.*, 2015] learns an unsupervised re-ranking model that removes the visual ambiguities by analyzing the content and context information in the initial ranking list. [Ye *et al.*, 2016] attempts to tackle the re-ranking problem by exploiting the common nearest neighbors. To address the false match issue from k-nearest neighbors, [Zhong *et al.*, 2017a] proposes to utilize k-reciprocal neighbors and designs an encoding method to revise the initial rank list by calculating feature distance and jaccard distance of samples.

Most existing re-ranking methods are designed for single-modality re-ID which do not work well in the cross-modality re-ID task. The major problem is that existing re-ranking methods cannot re-rank the samples of different modalities which have different similarity metrics as compared with samples of a single modality. We tackle this problem by combining cross-modality k-nearest neighbors and intra-modality k-reciprocal neighbors which improves the re-ID performance significantly, more details to be described in Sec.3.3.

3 The Proposed Approach

Given a query infrared person image $\boldsymbol{q}$ and the gallery set with N_g RGB images $\boldsymbol{\mathcal{G}} = \{\boldsymbol{g}_j \mid j = 1, 2, ..., N_g\}$, cross-modality re-ID ranks the gallery images according to their similarities to the query $\boldsymbol{q}$. Existing methods usually derive the similarity metric by directly comparing features of cross-modality samples, which often face low precision due to the gap and bias across image modalities. The proposed Similarity Inference Metric (SIM) aims to improve the cross-modality similarity metrics by exploiting the discriminative intra-modality similarity among gallery samples. It consists of feature representation, similarity graph reasoning, and mutual nearest-neighbor reasoning as illustrated in Fig. 2, more details to be described in the ensuing subsections.

3.1 Feature Representation

A weight-sharing two-stream CNN structure is designed to learn features and image representation from infrared and RGB images as illustrated in Fig. 2a. The CNN models are trained by optimizing cross entropy loss and triplet loss with infrared and RGB training samples together.

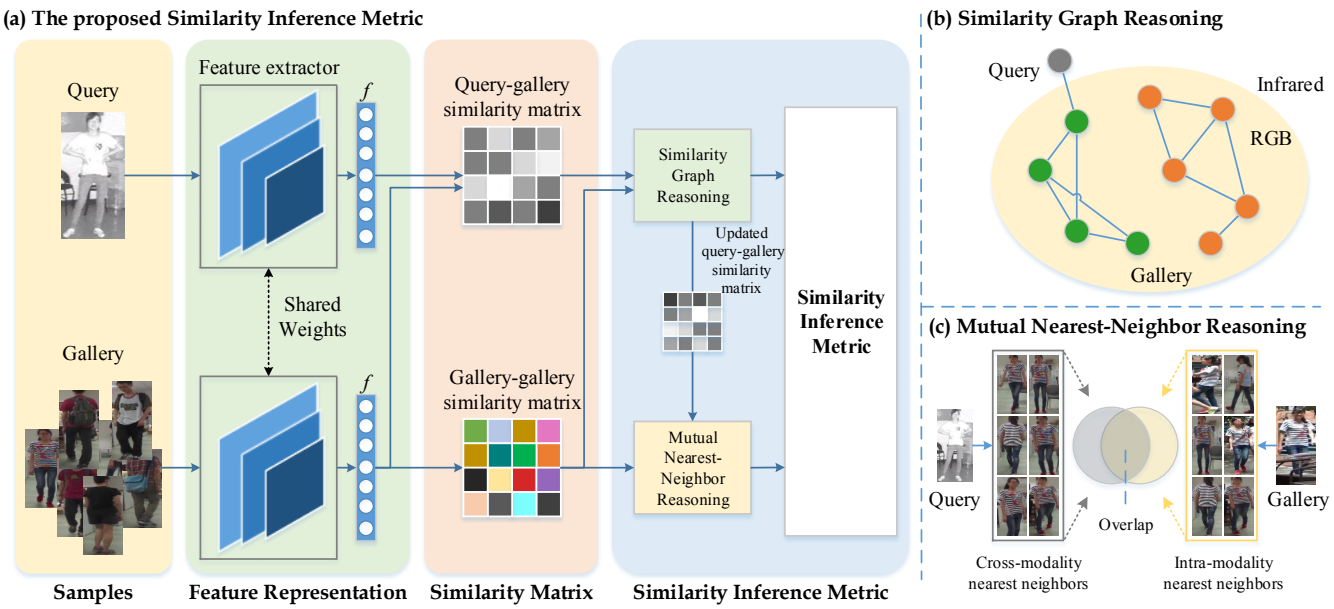

Figure 2: (a) The flowchart of the proposed similarity inference metric (SIM): SIM consists of two types of reasoning including similarity graph reasoning and mutual nearest-neighbor reasoning. (b) Similarity graph reasoning introduced in Sec. 3.2, where each color (green or orange) refers to samples of the same ID. (c) Mutual nearest-neighbor reasoning introduced in Sec. 3.3.

In inference phase, each infrared query image $\boldsymbol{q}_i$ in the query set $\boldsymbol{\mathcal{Q}} = \{\boldsymbol{q}_i \mid i = 1, 2, ..., N_q\}$ and each RGB gallery image $\boldsymbol{g}_j$ from $\boldsymbol{G}$ are fed to the trained model to extract respective features $\boldsymbol{f}_{\boldsymbol{q}_i}$ and $\boldsymbol{f}_{\boldsymbol{g}_j}$. A query-gallery similarity matrix $\boldsymbol{D}_{\boldsymbol{q},\boldsymbol{g}} \in \mathbb{R}^{N_q \times N_g}$ can then be obtained by computing L_2 distance between all query images and gallery images, where each matrix element $\boldsymbol{D}(i, j)$ denotes the distance between $\boldsymbol{f}_{\boldsymbol{q}_i}$ and $\boldsymbol{f}_{\boldsymbol{g}_j}$. Similarly, a gallery-gallery similarity matrix $\boldsymbol{D}_{\boldsymbol{g},\boldsymbol{g}} \in \mathbb{R}^{N_g \times N_g}$ can be obtained for all image pairs in the gallery set. Due to abundant optical information with little modality gap, $\boldsymbol{D}_{\boldsymbol{g},\boldsymbol{g}}$ is much more discriminative than $\boldsymbol{D}_{\boldsymbol{q},\boldsymbol{g}}$.

3.2 Similarity Graph Reasoning

We propose similarity graph reasoning to circumvent the cross-modality discrepancy by leveraging the intra-modality similarity. The idea is that for a query image $\boldsymbol{q}$ and its similar gallery image $\boldsymbol{g}$, other gallery images that are similar to $\boldsymbol{g}$ should be similar to $\boldsymbol{q}$ (via $\boldsymbol{g}$) even though they may have large distances from $\boldsymbol{q}$ as illustrated in Fig. 2b. With the matrices $\boldsymbol{D}_{\boldsymbol{q},\boldsymbol{g}}$ and $\boldsymbol{D}_{\boldsymbol{g},\boldsymbol{g}}$ as described in the previous subsection, we define a similarity graph $\mathbf{A}(\boldsymbol{\mathcal{V}}, \boldsymbol{\mathcal{E}})$ on the whole test set including all query and gallery, where each node in $\boldsymbol{\mathcal{V}} = \{\boldsymbol{\mathcal{Q}}; \boldsymbol{\mathcal{G}}\}$ represents an image sample and each edge in $\boldsymbol{\mathcal{E}}$ represents the similarity between its connected two nodes. We initialize the cross-modality edges (query-gallery) with $\boldsymbol{D}_{\boldsymbol{q},\boldsymbol{g}}$ and intra-modality edges (gallery-gallery) with $\boldsymbol{D}_{\boldsymbol{g},\boldsymbol{g}}$ as follows:

$$\begin{cases} \boldsymbol{\mathcal{E}}(\boldsymbol{q}_i, \boldsymbol{g}_j) = \boldsymbol{D}_{\boldsymbol{q},\boldsymbol{g}}(\boldsymbol{q}_i, \boldsymbol{g}_j) \\ \boldsymbol{\mathcal{E}}(\boldsymbol{g}_j, \boldsymbol{g}_k) = \lambda \boldsymbol{D}_{\boldsymbol{g},\boldsymbol{g}}(\boldsymbol{g}_j, \boldsymbol{g}_k), \end{cases} \tag{1}$$

where $\lambda \in [0, 1]$ is a scale factor that adjusts the ratio of two distance space.

Given the query image $\boldsymbol{q}_i$ and the gallery image $\boldsymbol{g}_j$, the distance $d(\boldsymbol{q}_i, \boldsymbol{g}_j)$ in perspective of similarity graph reasoning is defined as the shortest path from $\boldsymbol{q}_i$ to $\boldsymbol{g}_j$ between the query node $\boldsymbol{q}_i$ and the gallery nodes $\boldsymbol{\mathcal{G}}$ on the Graph $\mathbf{A}$. To be specific, suppose $\boldsymbol{\Omega}_{\boldsymbol{q}_i,\boldsymbol{g}_j}$ denotes the set that includes all the possible path from $\boldsymbol{q}_i$ to $\boldsymbol{g}_j$. For any path $\mathcal{P} \in \boldsymbol{\Omega}_{\boldsymbol{q}_i,\boldsymbol{g}_j}$, $\mathcal{P} = (\boldsymbol{p}_1, \boldsymbol{p}_2, ..., \boldsymbol{p}_n)$ where $n = length(\mathcal{P}), \boldsymbol{p}_1 = \boldsymbol{q}_i, \boldsymbol{p}_n = \boldsymbol{g}_j, \boldsymbol{p}_k \in \boldsymbol{\mathcal{G}}$ $(2 \leq k \leq n-1)$, $d(\boldsymbol{q}_i, \boldsymbol{g}_j)$ is formulated as

$$d(\boldsymbol{q}_i, \boldsymbol{g}_j) = \min_{\mathcal{P} \in \boldsymbol{\Omega}_{\boldsymbol{q}_i,\boldsymbol{g}_j}} \sum_{t=1}^{n-1} \boldsymbol{\mathcal{E}}(\boldsymbol{p}_t, \boldsymbol{p}_{t+1}). \tag{2}$$

Due to the fact that L_2 metric used in gallery satisfies triangle inequality below

$$\boldsymbol{\mathcal{E}}(\boldsymbol{g}_i, \boldsymbol{g}_j) + \boldsymbol{\mathcal{E}}(\boldsymbol{g}_j, \boldsymbol{g}_t) \geq \boldsymbol{\mathcal{E}}(\boldsymbol{g}_i, \boldsymbol{g}_t), \quad \forall\, 1 \leq i, j, t \leq N_g. \tag{3}$$

Thus, the query-gallery distance can be simplified by:

$$d(\boldsymbol{q}_i, \boldsymbol{g}_j) = \min_{1 \leq t \leq N_g} \{\boldsymbol{\mathcal{E}}(\boldsymbol{q}_i, \boldsymbol{g}_t) + \boldsymbol{\mathcal{E}}(\boldsymbol{g}_t, \boldsymbol{g}_j)\}. \tag{4}$$

Further, we use the mean of the first K shortest paths instead of the shortest one for more stable cross-modality distance evaluation as follows:

$$d_S(\boldsymbol{q}_i, \boldsymbol{g}_j) = \frac{1}{K} \sum_{k=1}^{K} d^{(k)}(\boldsymbol{q}_i, \boldsymbol{g}_j). \tag{5}$$

where $d^{(k)}(\boldsymbol{q}_i, \boldsymbol{g}_j)$ denotes the k-th shortest path between $\boldsymbol{q}_i$ and $\boldsymbol{g}_j$.

In practice, to reduce the computational complexity, all useless edges between gallery pairs are deleted except those between each sample and its k-nearest neighbors in the gallery set.

Algorithm 1 Similarity Inference Metric (SIM)

Input: Query-gallery similarity matrix $\boldsymbol{D_{q,g}}$, gallery-gallery similarity matrix $\boldsymbol{D_{g,g}}$
Parameter: λ, K and α
Output: SIM $\boldsymbol{d}_{SIM}$

Initialize similarity graph $\mathbf{A}(\boldsymbol{\mathcal{V}}, \boldsymbol{\mathcal{E}})$ as Eq. (1).
%*Compute SGR distance*
for each $\boldsymbol{q}_i$, $\boldsymbol{g}_j$ **do**
  for $\boldsymbol{g}_t$ in $\boldsymbol{g}_j$'s k-nearest neighbors **do**
    Calculate $d(\boldsymbol{q}_i, \boldsymbol{g}_t, \boldsymbol{g}_j) = \boldsymbol{\mathcal{E}}(\boldsymbol{q}_i, \boldsymbol{g}_t) + \boldsymbol{\mathcal{E}}(\boldsymbol{g}_t, \boldsymbol{g}_j)$
  end for
  Sort $d(\boldsymbol{q}_i, \boldsymbol{g}_t, \boldsymbol{g}_j)$ for all $\boldsymbol{g}_t$.
  Calculate $\boldsymbol{d}_S(\boldsymbol{q}_i, \boldsymbol{g}_j)$ according to Eq. (5).
end for
%*Compute MNNR distance*
for each $\boldsymbol{q}_i$, $\boldsymbol{g}_j$ **do**
  Calculate $\boldsymbol{d}_M(\boldsymbol{q}_i, \boldsymbol{g}_j)$ according to Eq. (7).
end for
Calculate $\boldsymbol{d}_{SIM}$ according to Eq. (8).
return $\boldsymbol{d}_{SIM}$

3.3 Mutual Nearest-Neighbor Reasoning

We proposed mutual nearest-neighbor reasoning (MNNR) under the hypothesis that a query image $\boldsymbol{q}_i$ and a gallery image $\boldsymbol{g}_j$ are more likely to be a true match if they have the same mutual k-nearest neighbors in the gallery set. Neighbor information has been explored in re-ranking based re-ID, e.g. by k-reciprocal encoding [Zhong *et al.*, 2017a]. But it was mainly used for single-modality re-ID which does not work well in cross-modality re-ID where similarity metrics of query-gallery and gallery-gallery are discrepant. For example, the cross-modality query-gallery distance $D_{\boldsymbol{q},\boldsymbol{g}}$ and the intra-modality gallery-gallery distance $D_{\boldsymbol{g},\boldsymbol{g}}$ are often at different scales and cannot be ranked while handling test samples of different modalities.

MNNR employs a series of asymmetric strategies to handle the cross-modality discrepancy as shown in Fig. 2c. First, it uses gallery set as the search space without including query. For an IR query $\boldsymbol{q}$, it ranks gallery images with similarities d_S and obtains its $k_{\boldsymbol{q}}$ cross-modality nearest neighbors:

$$\boldsymbol{\mathcal{N}}_c(\boldsymbol{q}_i, k_{\boldsymbol{q}}, d_S) = \{\boldsymbol{g}^{(1)}, \boldsymbol{g}^{(2)}, ..., \boldsymbol{g}^{(k_q)}\}. \tag{6}$$

For a RGB gallery image $\boldsymbol{g}$, it ranks the gallery images with $\boldsymbol{D_{g,g}}$ and obtains its $k_{\boldsymbol{g}}$ intra-modality reciprocal nearest neighbors as $\boldsymbol{\mathcal{R}}_i^*(\boldsymbol{g}, k_{\boldsymbol{g}}, \boldsymbol{D_{g,g}})$. The mutual nearest neighbors of $\boldsymbol{q}$ and $\boldsymbol{g}$ can thus be defined by the overlap between $\boldsymbol{\mathcal{N}}_c(\boldsymbol{q}, k_{\boldsymbol{q}}, d_S)$ and $\boldsymbol{\mathcal{R}}_i^*(\boldsymbol{g}, k_{\boldsymbol{g}}, \boldsymbol{D_{g,g}})$. Intuitively, more mutual nearest neighbors means higher similarity and the MNNR distance d_M can be defined by:

$$d_M(\boldsymbol{q}_i, \boldsymbol{g}_j) = 1 - \frac{|\boldsymbol{\mathcal{N}}_{\boldsymbol{c}}(\boldsymbol{q}_i, k_{\boldsymbol{q}}, d_S) \cap \boldsymbol{\mathcal{R}}_i^*(\boldsymbol{g}_j, k_{\boldsymbol{g}}, \boldsymbol{D_{g,g}})|}{|\boldsymbol{\mathcal{N}}_{\boldsymbol{c}}(\boldsymbol{q}_i, k_{\boldsymbol{q}}, d_S) \cup \boldsymbol{\mathcal{R}}_i^*(\boldsymbol{g}_j, k_{\boldsymbol{g}}, \boldsymbol{D_{g,g}})|}. \tag{7}$$

where $|\cdot|$ denotes the number of candidates in the set.

3.4 Similarity Inference Metric

The proposed Similarity Inference Metric can thus be derived by combining the similarity graph reasoning and mutual nearest-neighbor reasoning. It jointly aggregates d_S and d_M as the final distance as follows:

$$\boldsymbol{d}_{SIM} = \alpha \boldsymbol{d}_S + (1 - \alpha)\boldsymbol{d}_M. \tag{8}$$

where $\alpha \in [0, 1]$ denotes the penalty factor. When $\alpha = 1$, only the similarity graph reasoning is considered. **Algorithm 1** provides the detailed description of our proposed similarity inference metric.

3.5 Complexity Analysis

Most of computations focus on pairwise distance calculation and distance ranking for all gallery pairs and the computation complexity is $O(N_g^2)$ and $O(N_g^2 \log N_g)$, respectively. Given a new query $\boldsymbol{q}$, SIM just computes the distance between $\boldsymbol{q}$ and gallery ($O(N_g)$), ranks all path distance for SGR ($O(K \log K)$), computes the $k_{\boldsymbol{q}}$-nearest neighbors for MNNR ($O(N_g \log N_g)$), and ranks the final distance ($O(N_g \log N_g)$).

4 Experiments

4.1 Datasets and Settings

The proposed SIM is evaluated over two public datasets RegDB and SYSU-MM01. The standard Cumulative Matching Characteristic (CMC) curve and mean average precision (mAP) are adopted as the evaluation metrics. Different from the traditional single-modality re-ID, the evaluations here use IR images as probe set and RGB images as gallery set for both datasets.

RegDB [Nguyen *et al.*, 2017] is collected by using dual cameras (with optical and thermal sensors). It contains images of 412 persons, where for each person 10 RGB images and 10 IR images are collected. Following the evaluation protocol in (Ye et al.2018), this dataset is randomly split into two halves, one half for training and the other half for testing. In addition, the evaluation procedure is repeated for 10 trials to achieve statistically stable results.

SYSU-MM01 [Wu *et al.*, 2017] is a large-scale RGB-IR re-ID dataset which contains images of 419 identities captured using six disjoint surveillance cameras (four RGB cameras and two IR cameras). The training set contains 19,659 RGB images and 12,792 IR images of 395 persons and the test set contains images of 96 persons. Following [Wu *et al.*, 2017], we adopt the multi-shot *all-search* mode evaluation protocol where 10 images of a person are randomly selected to form the gallery set with 10 times repeat.

Implementation details. We adopt the ResNet-50 [He *et al.*, 2016] as the backbone network and initialize it by using parameters pre-trained on the ImageNet [Krizhevsky *et al.*, 2012]. During training, the input image is uniformly resized to 256×128 and traditional image augmentation is performed via random flipping and random erasing [Zhong *et al.*, 2017b]. In addition, we use the Adam optimizer to train the model and the learning rate is set at 3.5×10^{-4}. The whole training process consists of 200 epochs.

4.2 Comparison with State-of-the-Arts

The proposed SIM is compared with a number of cross-modality person re-ID methods that can be broadly classified into three categories: 1) LOMO [Liao *et al.*, 2015],

Methods	Visible2Thermal		Thermal2Visible	
	mAP	Rank-1	mAP	Rank-1
LOMO	2.28	0.85	-	-
HOG	10.31	13.49	-	-
GSM	15.06	17.28	-	-
One-stream	14.02	13.11	-	-
Two-stream	12.43	30.36	-	-
Zero-Padding	18.90	17.75	17.82	16.63
TONE	14.92	16.87	16.98	13.86
HCML	20.08	24.44	22.24	21.70
BDTR	31.83	33.47	31.10	32.72
D-HSME	47.0	50.9	46.2	50.2
D^2RL	44.1	43.4	44.1	43.4
PIG	49.3	48.5	48.9	48.1
AlignGAN	53.60	57.90	53.40	56.30
SIM (ours)	**75.29**	**74.47**	**78.30**	**75.24**

Table 1: Comparison with state-of-the-art cross-modality re-ID methods over the dataset RegDB: Visible2Thermal means using RGB images as query and IR images as gallery, and Thermal2Visible means the opposite.

Methods	mAP	Rank-1
LOMO	2.28	4.70
HOG	2.16	3.82
GSM	4.38	6.19
One-stream	8.59	16.3
Two-stream	8.03	16.4
Zero-Padding	10.9	19.2
cmGAN	22.27	31.49
PIG	29.5	45.1
AlignGAN	33.90	51.50
SIM (ours)	**60.88**	**56.93**

Table 2: Comparison with state-of-the-art cross-modality re-ID methods over the dataset SYSU-MM01

HOG [Dalal and Triggs, 2005] and GSM [Lin *et al.*, 2016] that use hand-crafted features; 2) One-stream, Two-stream, Zero-Padding [Wu *et al.*, 2017], TONE [Ye *et al.*, 2018a], BDTR [Ye *et al.*, 2018b] and cmGAN [Dai *et al.*, 2018] that focus on feature distribution alignment; and 3) D^2RL [Wang *et al.*, 2019b], PIG [Wang *et al.*, 2020] and AlignGAN [Wang *et al.*, 2019a] that use GANs to transfer image styles. Table 1 and Table 2 show the experimental results over the datasets RegDB and SYSU-MM01, respectively, where Visible2Thermal means using RGB images as query and IR images as gallery, and Thermal2Visible means the opposite.

As the two tables show, methods in the first category do not perform well due to the constraints of hand-crafted features. Methods in the second category learn modality-invariance features by suppressing feature distribution gaps across modality, which achieve clearly better re-ID performance. GAN based methods reduce the modality discrepancy at image level which further improve the re-ID performance.

Our proposed Similarity Inference Metric outperforms all the competing methods significantly. As Table 1 shows, it outperforms the state-of-the-art (AlignGAN) by 24.9% in mAP (78.3% vs 53.4%) and 18.94% in rank-1 accuracy (75.24% vs 56.3%) for Thermal2Visible. Similar improvement is obtained for the Visible2Thermal. For the dataset SYSU-MM01, SIM obtains an mAP of 60.88% and a rank-1 of 56.93% as shown in Table 2, which outperforms the state-of-the-art (AlignGAN) by 26.98% and 5.43%, respectively.

Methods	SGR	MNNR	mAP	Rank-1
Baseline	×	×	39.90	53.93
SGR only	✓	×	59.17	56.62
MNNR only	×	✓	54.39	56.31
SIM	✓	✓	60.88	56.93

Table 3: Ablation studies of our proposed Similarity Inference Metric over SYSU-MM01: *Baseline* uses the traditional metric, i.e. L_2 distance between image features; *SGR only* incorporates the Similarity Graph Reasoning (SGR) only over the *Baseline*; *MNNR only* incorporates the Mutual Nearest-Neighbor Reasoning (MNNR) only over the *Baseline*; *SIM* incorporates both SGR and MNNR.

4.3 Ablation Studies

Extensive ablation studies have been performed to evaluate each component of our proposed SIM. As Table 3 shows, four networks are trained including: 1) *Baseline* that uses the traditional L_2 distance to measure the feature similarity; 2) *SGR only* that just incorporates the Similarity Graph Reasoning (as described in Section 3.2) beyond the *Baseline*; 3) *MNNR only* that just incorporates the proposed Mutual Nearest-Neighbor Reasoning (as described in Section 3.3) beyond the *Baseline*; and *SIM* that incorporates both SGR and MNNR. As Table 3 shows, the *Baseline* does not perform well due to the large discrepancy across image modalities.

In addition, either *SGR only* or *MNNR only* improves the re-ID performance greatly. Specifically, *SGR only* achieves a mAP of 59.17% and a rank-1 accuracy of 56.62%, which are higher than the *Baseline* by 19.27% and 2.69%, respectively. This results show that SGR improves the sample similarity greatly by exploiting the discriminative within-gallery similarities. Similarly, *MNNR only* improves the mAP by 14.49% and the rank-1 accuracy by 2.37%, respectively, as compared with the *Baseline*. The effectiveness of the MNNR can be largely attributed to the use of the overlap of k-nearest neighbor sets in gallery between image pairs.

Further, SIM with both SGR and MNNR outperforms either *SGR only* or *MNNR only*, demonstrating the complementariness of the two proposed reasoning mechanisms. It achieves a mAP of 60.88% and a rank-1 accuracy of 56.93% which are higher than the *Baseline* by 20.98% and 3.00%, respectively. This shows that the proposed SIM enhances the cross-modality sample similarities effectively. The contribution of our SIM can also be observed in the ranking list as illuminated in Fig. 3. SIM effectively improves the similarities of true persons which are ranked behind of the baseline.

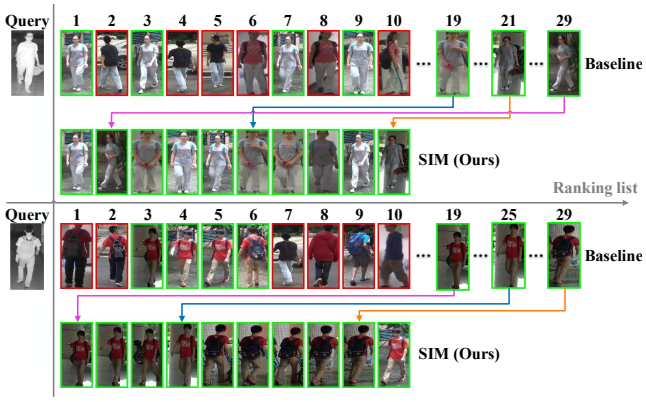

Figure 3: Illustration of how SIM helps improve cross-modality re-ID on the SYSU-MM01 dataset: With SIM, the similarity of IR and RGB images is improved greatly – the three lines in blue, orange, and pink show three examples of improved image similarities by our proposed SIM. Person images in red (green) boxes denote the negative (positive) samples.

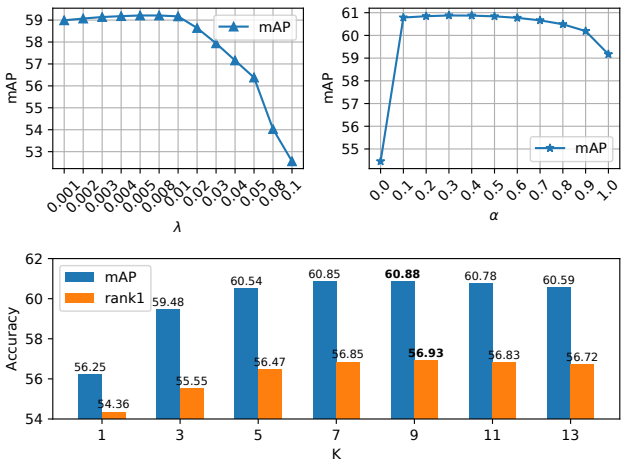

Figure 4: The impact of the parameter λ, α and K on re-ID performance on the SYSU-MM01 dataset.

4.4 Parameters Analysis

The proposed SIM involves three key parameters including scale factor λ, limit factor K and penalty factor α. The three parameters are studied by setting them to different values and checking the corresponding re-ID performance as shown in Fig. 4. As the top-left graph shows, λ should be small so as to exploit the within-gallery sample similarity sufficiently (α and K fixed at 0.01 and 9). This can also be observed for α and K. For example, when K is set at a small value 1, the false positive matching increases clearly as it lowers the fault tolerance for the first matching. On the contrary, the re-ID performance is degraded due to its weak discrimination when K is set at 13. Experiments show that SIM performs optimally when $\lambda = 0.01$, $K = 9$ and $\alpha = 0.3$.

4.5 Generalization Analysis

The proposed SIM is a generic metric that can work with different existing re-ID methods. We study this nice property by applying SIM to AlignGAN [Wang *et al.*, 2019a] and AGW [Ye *et al.*, 2020], both using the traditional L_2 distance metric in feature similarity evaluation. Table. 4 shows experimental results. As Table. 4 shows, either SGR or MNNR improves the re-ID performance by large margins when it is incorporated into the AlignGAN and AGW methods. In addition, further improvements are observed when both SGR and MNNR are incorporated. These results are well aligned with the experimental results observed in the Ablation Studies.

Specifically, AlignGAN* + SGR achieves a mAP of 51.30% and a rank-1 accuracy of 52.56% which are higher than the AlignGAN by 16.74% and 4.64%, respectively. AlignGAN* + MNNR achieves a mAP of 50.26% and a rank-1 accuracy of 52.33% which also outperforms AlignGAN significantly. Similar improvement can be observed for AGW as well. All these experimental results demonstrate the superiority of our proposed Similarity Inference Metric (with SGR and MNNR) that can generalize across different cross-modality re-ID metrics with significant and consistent performance improvements but little extra training.

Methods	mAP	Rank-1
AlignGAN*[Wang *et al.*, 2019a]	34.56	47.92
AlignGAN* + SGR	51.30	52.56
AlignGAN* + MNNR	50.26	52.33
AlignGAN* + SIM (SGR+MNNR)	**54.45**	**52.70**
AGW [Ye *et al.*, 2020]	40.03	50.87
AGW + SGR	55.89	52.70
AGW + MNNR	51.40	52.93
AGW + SIM (SGR+MNNR)	**57.47**	**53.75**

Table 4: Generalization Analysis of Similarity Inference Metric on SYSU-MM01 dataset with other approach as baseline, *i.e.* AlignGAN. AlignGAN* denotes our re-implemented version.

5 Conclusion

This paper presents an innovative Similarity Inference Metric (SIM) for RGB-Infrared person re-identification. We introduce similarity graph reasoning and mutual nearest-neighbor reasoning to infer inter-modality sample similarities by exploiting reliable intra-modality sample similarity. The two types of reasoning can generalize over different cross-modality person re-ID metrics with significant performance improvements but little extra training. Experiments demonstrate the effectiveness as well as the generalization of our method for improving re-ID performance. We expect that the proposed SIM will inspire new insights for better cross-modality person re-ID in the near future.

Acknowledgements

This work was supported in part by National Natural Science Foundation of China (61902009) and Shenzhen Research Project (201806080921419290).

References

[Chen *et al.*, 2017] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *CVPR*, pages 403–412, 2017.

[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*, pages 677–683, 2018.

[Dalal and Triggs, 2005] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[Garcia *et al.*, 2015] Jorge Garcia, Niki Martinel, Christian Micheloni, and Alfredo Gardel. Person re-identification ranking optimisation by discriminant context information analysis. In *ICCV*, pages 1305–1313, 2015.

[Gou *et al.*, 2014] Mengran Gou, Xiong Fei, Octavia Camps, and Mario Sznaier. Person re-identification using kernel-based metric learning methods. In *ECCV*, 2014.

[Hao *et al.*, 2019] 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*, pages 770–778, 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.

[Krizhevsky *et al.*, 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, pages 1097–1105, 2012.

[Li *et al.*, 2018] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *CVPR*, pages 2285–2294, 2018.

[Liao *et al.*, 2015] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, pages 2197–2206, 2015.

[Lin *et al.*, 2016] Liang Lin, Guangrun Wang, Wangmeng Zuo, Xiangchu Feng, and Lei Zhang. Cross-domain visual matching via generalized similarity measure and feature learning. *TPAMI*, 39(6):1089–1102, 2016.

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

[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*, October 2019.

[Wang *et al.*, 2019b] 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*, pages 618–626, 2019.

[Wang *et al.*, 2020] Guan'an Wang, Tianzhu Zhang, Yang Yang, Jian Cheng, jianlong Chang, Xu Liang, and Zengguang Hou. Cross-modality paired-images generation for rgb-infrared person re-identification. In *AAAI*, February 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*, pages 5380–5389, 2017.

[Yang *et al.*, 2019] Fan Yang, Ke Yan, Shijian Lu, Huizhu Jia, Xiaodong Xie, and Wen Gao. Attention driven person re-identification. *Pattern Recognit.*, 86:143–155, 2019.

[Ye *et al.*, 2016] Mang Ye, Chao Liang, Yi Yu, Zheng Wang, Qingming Leng, Chunxia Xiao, Jun Chen, and Ruimin Hu. Person reidentification via ranking aggregation of similarity pulling and dissimilarity pushing. *IEEE Transactions on Multimedia*, 18(12):2553–2566, 2016.

[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, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *arXiv preprint arXiv:2001.04193*, 2020.

[Zhai *et al.*, 2020] Yunpeng Zhai, Shijian Lu, Qixiang Ye, Xuebo Shan, Jie Chen, Rongrong Ji, and Yonghong Tian. Ad-cluster: Augmented discriminative clustering for domain adaptive person re-identification. In *IEEE CVPR*, 2020.

[Zhao *et al.*, 2017] Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, and Xiaoou Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *CVPR*, pages 1077–1085, 2017.

[Zhen *et al.*, 2013] Li Zhen, Shiyu Chang, Liang Feng, Thomas S. Huang, and John R. Smith. Learning locally-adaptive decision functions for person verification. In *CVPR*, 2013.

[Zheng *et al.*, 2011] Wei Shi Zheng, Shaogang Gong, and Xiang Tao. Person re-identification by probabilistic relative distance comparison. In *CVPR*, 2011.

[Zhong *et al.*, 2017a] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *CVPR*, July 2017.

[Zhong *et al.*, 2017b] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. *arXiv preprint arXiv:1708.04896*, 2017.